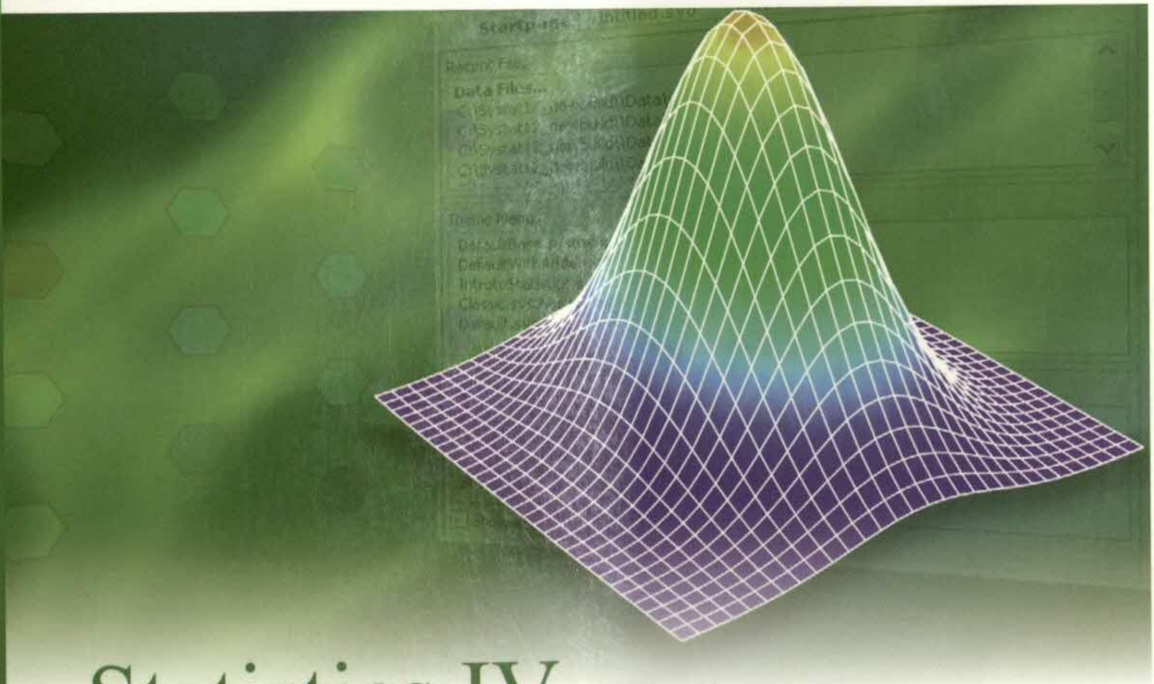
337

7

SYSTAT® 12



Statistics IV

For Acc
SCERS Lib

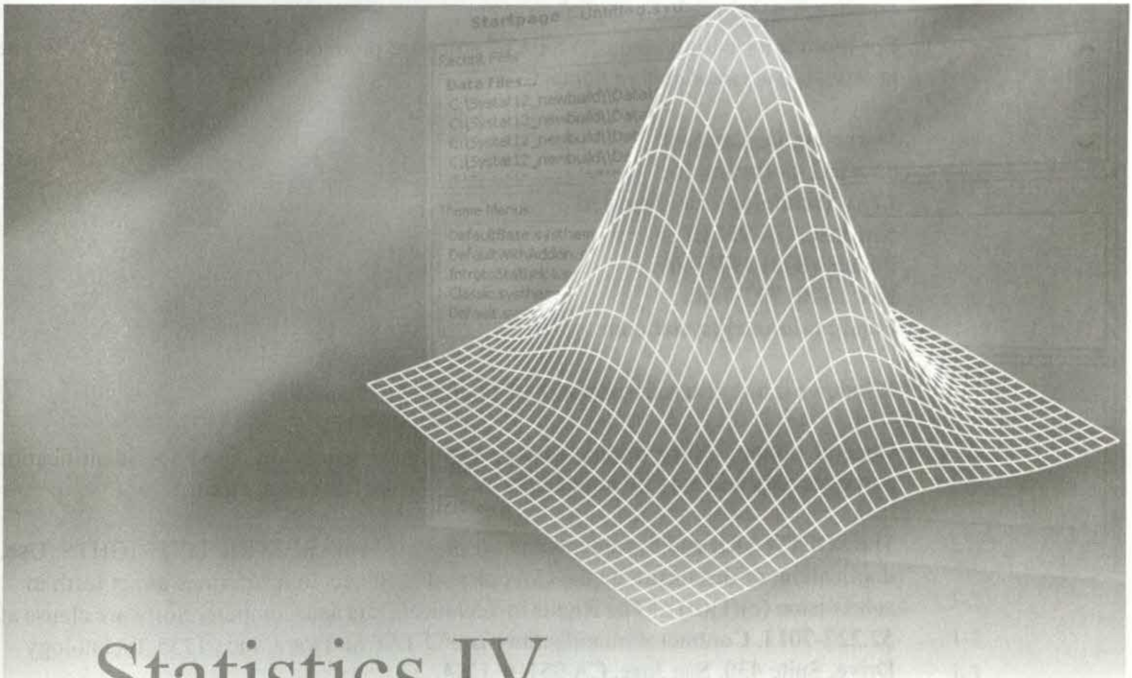
28/5/10

SYSTAT®

WWW.SYSTAT.COM

333

SYSTAT® 12



Statistics IV

SYSTAT

WWW.SYSTAT.COM

For more information about SYSTAT® software products, please visit our WWW site at *http://www.systat.com* or contact

Marketing Department
SYSTAT Software, Inc.
1735 Technology Dr., Ste. 430
San Jose, CA 95110
Phone: (800) 797-7401
Fax: (800) 797-7406
Email: info-usa@systat.com

SYSTAT® 12 Statistics- IV

SYSTAT Software, Inc.
1735 Technology Dr., Ste. 430
San Jose, CA 95110

Printed in the United States of America.

1 2 3 4 5 6 7 8 9 0   05 04 03 02 01 00

Contents

2 Bootstrapping and Sampling I-17

3 Classification and Regression Trees I-41

4 Cluster Analysis *I-65*

5 Conjoint Analysis *I-125*

6 Correlations, Associations, and Distance Measures I-157

7 Correspondence Analysis I-201

10 Design of Experiments *I-345*

11 Discriminant Analysis I-391

12 Factor Analysis I-453

13 Fitting Distributions *I-495*

14 Hypothesis Testing *I-519*

Statistics II

1 Linear Models II-1

6 Variance Components Models II-299

7 Linear Mixed Models II-343

8 Hierarchical Linear Mixed Models *II-385*

9 Mixed Regression *II-421*

Statistics III

2 Loglinear Models *III-93*

3 Missing Value Analysis *III-123*

7 Nonlinear Models *III-261*

8 Nonparametric Tests *III-319*

10 Partially Ordered Scalogram Analysis with Coordinates *III-381*

11 Path Analysis (RAMONA) *III-397*

Statistics IV

3 Probability Calculator *IV-85*

4 Probit Analysis IV-99

5 Quality Analysis IV-109

6 Random Sampling *IV-219*

7 Response Surface Methods IV-231

8 Robust Regression IV-255

9 Set and Canonical Correlations IV-291

10 Signal Detection Analysis IV-319

11 Smoothing IV-349

12 Spatial Statistics *IV-385*

13 Survival Analysis *IV-427*

16 Two-Stage Least Squares *IV-581*

Acronym & Abbreviation Expansions

Index

List of Examples

Chapter

1

Perceptual Mapping

Leland Wilkinson

PERMAP offers two types of tools. The first is a group of procedures for fitting subjects and objects in a common space. This group includes Carroll's (1972) internal and external unfolding models, MDPREF and PREFMAP, as well as Gabriel's (1971) BIPLOT, which is a minor modification of MDPREF. The second is a set of procedures for relating one dimensional configuration to another, generally called a procrustes rotation. Both the orthogonal procrustes and the more general canonical rotations are available.

PERMAP is a misnomer. Although most of the techniques it incorporates have been used for perceptual mapping, they have applications outside of market research or psychology and, like the biplot technique, may even have their origins elsewhere. Furthermore, classical perceptual mapping techniques, such as multidimensional scaling, correspondence analysis, and principal components, are found elsewhere in SYSTAT. In the end, since almost all of the methods in this module involve a singular value decomposition and are not bulky enough to deserve their own modules, they have been collected into a single grab bag.

Statistical Background

Perceptual mapping involves a variety of techniques for displaying the judgments of a set of objects by a group of subjects. Most of these techniques were developed in the 1970's by psychometricians, but they were soon adopted by market researchers and scientists for analyzing a variety of preference and similarity data.

In applied usage, especially among market researchers, perceptual mapping is an even more general term. Some commercial perceptual mapping programs are based

on classical statistical or psychometric models. Some of these methods include Fisher's linear discriminant function, correspondence analysis, factor analysis, and multidimensional scaling. Indeed, any procedure that produces a set of coordinates in a q dimensional space from an $n \times p$ matrix, $q \le \min(n, p)$, can be considered perceptual mapping in the broad, applied sense. Quantitative theoretical market researchers (for example, Green and Tull, 1975, and Lilien, Kotler, and Moorthy, 1992) use the term in this more general sense as well.

The origin of the term can be found in classical psychometrics (see Cliff, 1973, for a review). Soon after the development of psychometric spatial models, some psychologists thought scaling methods could be used to derive "cognitive maps" from the subject's ratings of stimuli. These maps would be "pictures" of the mental structures used to perceive and integrate information. Following the classic linguistic studies of Osgood, Suci, and Tannenbaum (1957), researchers produced intriguing cognitive maps of stimuli such as countries, cities, adjectives, colors, and consumer products (for example, Wish, Deutsch, and Biener, 1972, and Milgram and Jodelet, 1976).

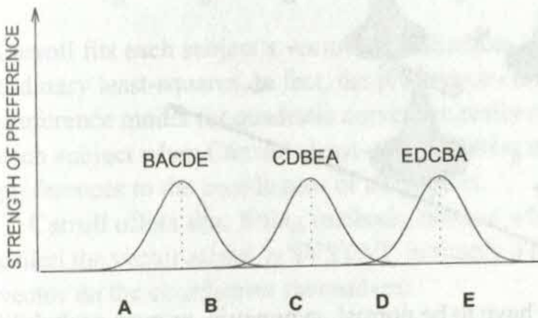
Not long afterwards, perceptual and memory psychologists abandoned the cognitive map model and developed theories based on information processing, problem solving, and associative memory. Research by Shepard and Cooper (1982) and Kosslyn (1981), for example, focused specifically on the storage and processing of mental images rather than inferring spatial structure among nonspatial stimuli from associations between responses to attributes. Shepard's psychometric findings on mental rotations, for example, were subsequently confirmed at the physiological level (Dow, 1990).

While no longer an active theoretical model, perceptual mapping can be useful as a general collection of procedures for presenting statistical analyses to nontechnical clients. Like classification trees, perceptual maps can show complex relations relatively simply without algebra or statistical parameters. It is easier for many clients to judge a distance on a map than to evaluate a conditional probability. Thus, perceptual mapping techniques can be useful for data that have nothing to do with perception.

Preference Mapping

A variety of algebraic and geometric models of preferences have been developed over the last century. The unidimensional preference model (Coombs, 1950) is presented in the following figure. Imagine that three subjects have expressed their preferences for each of five objects (*A, B, C, D, E*). If their preferences can be represented by a single

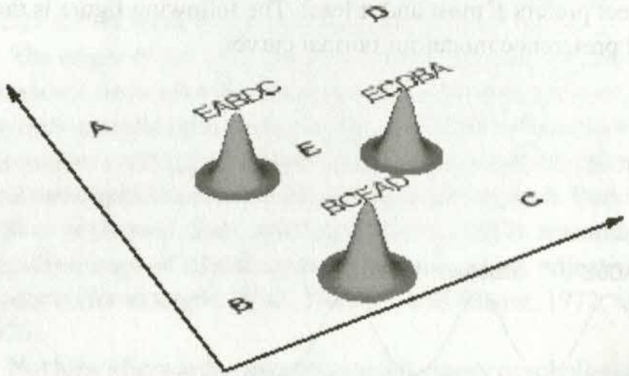
dimension, the following figure is one of several possible models. Each subject's preference strength on the single attribute dimension is represented by a normal distribution. In this model, the farther an object is from the mean of the subject's preference distribution, the less that object is preferred. Based on this rule, in the following figure, the ordering of preferences for the five objects shows above each subject's curve. Thus, the left most subject prefers object *B* most and *E* least, while the right most subject prefers *E* most and *A* least. The following figure is the unidimensional preference model for normal curves.



Coombs devised a method for recovering a unidimensional preference scale from the subject's ranking of the objects. His procedure is called *unfolding*. If we assume that the distances to objects from a subject's ideal point on the scale are all positive and follow the usual distance axioms (see Chapter "Multidimensional Scaling" on page 185 in *Statistics III*), then direction does not enter into the calculation of preference. We can therefore imagine folding the scale about the subject's ideal point to see the point of view of that subject. Coombs discovered that if there are enough subjects and objects, we can *unfold* the scale from the given ordering of the subject's preferences without knowing the strengths of the preferences. In general, the more the subjects and objects, the less room there is to represent the preference ordering correctly by moving the locations of the preference curves. Like MDS itself, the system becomes highly constrained to allow only one solution. The MDS procedure in SYSTAT can be used to compute Coombs's solution for unidimensional data.

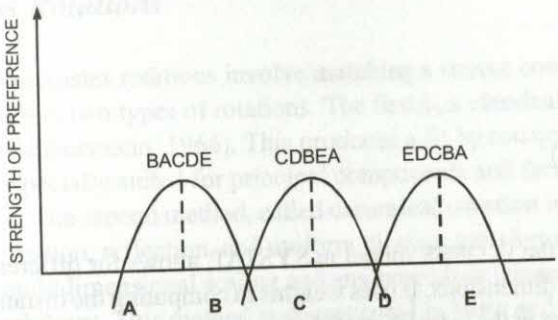
Students of Coombs (Bennet and Hays, 1960) extended the unfolding model to higher dimensions. The multidimensional preference model for normal curves shows how this works. As in the unidimensional preference model for normal curves, there are three subjects and five objects. The closest subject's preference curve leads to preferences of *BCEAD*, in left-to-right order. The other two subjects have preference curves nearest object *E* in the center of the configuration. Consequently, their most

preferred object is *E*. In the multidimensional model, distance is calculated in all directions from the center of the subject's preference curve. Again, the SYSTAT MDS module can be used to find solutions for multidimensional unfolding problems. The following figure is the multidimensional preference model for normal curves.



Preference curves do not have to be normal, symmetric, or even probability distributions. Carroll (1972) devised an unfolding procedure based on a quadratic preference curve model. He called the procedure "external" because it relied on quantitative ratings of the subject's preferences and a previously determined fixed configuration of objects in a space. While ordinary unfolding begins with an $n \times p$ matrix of *n* subjects' rank ordering of *p* objects, external unfolding begins with a $p \times q$ matrix of *p* objects' coordinates in *q* dimensions and a $p \times n$ matrix of *n* subjects' ratings of their preferences for the *p* objects.

The unidimensional preference model for quadratic curves shows Carroll's model in one dimension. Unlike the normal curve model, the quadratic preference curves involve negative preferences. The subject on the left in the unidimensional preference model for normal curves, for example, is indifferent about object *E* (or more indifferent than she is about object *D*). The subject on the left in the unidimensional preference model for quadratic curves, on the other hand, likes object *E* least. Carroll's model is therefore appropriate for data following a bipolar (approach-avoidance) preference model. The following is the unidimensional preference model for quadratic curves.



Carroll fits each subject's vector of preferences to a configuration of objects via ordinary least-squares. In fact, the preference curves in the previous unidimensional preference model for quadratic curves are really inverted (negative) quadratic loss for each subject when Carroll's least-squares fitting method is used to fit her vector of preferences to the coordinates of the objects.

Carroll offers four fitting methods, three of which appear in SYSTAT. The first, called the vector model in SYSTAT, is simply a multiple regression of the preference vector on the coordinates themselves:

$$s_{ij} = \sum_{k=1}^{q} a_{ik} x_{jk} + b_i + e_{ij}$$

where s_{ij} is the preference scale value of the *j*th stimulus for the *i*th subject. The coefficients a_{ik} are estimated by regressing ***y*** (the vector of preferences) on ***X*** (the $p \times q$ matrix of coordinates) and then transforming the coefficients.

This is called a vector model because the resulting fit is displayed as a vector superimposed on the object configuration. Preferences are predicted from the perpendicular projections of the object's coordinates onto each vector.

The second model, called the CIRCLE model in SYSTAT, is the one in the figures above. It is fit by regressing each subject's preferences on the coordinates and squared coordinates of the object configuration. From the coefficients in this regression, the ideal points are established in the coordinate space of the object configuration. In two dimensions, the intersection of each preference surface with the zero preference plane is a circle. The basic algebraic model is

$$s_{ij} = a_i d^2_{ij} + \sum_{k=1}^{q} b_{ik} x_{jk} + c_i + e_{ij}$$

where

$$d_{ij}^2 = \sum_{k=1}^{q} (x_{jk} - y_{ik})^2$$

The third model, called the ELLIPSE model in SYSTAT, allows for differential weighting of preference dimensions. It uses weights in computing the distances instead of the ordinary regression in the circular model. As a result, each preference curve may be elliptical at the zero preference plane. The model is

$$s_{ij} = a_i d_{ij}^2 + \sum_{k=1}^{q} b_{ik} x_{jk} + c_i + e_{ij}$$

where

$$d_{ij}^2 = \sum_{k=1}^{q} w_{ik}(x_{jk} - y_{ik})^2$$

Biplots and MDPREF

The CORAN procedure in SYSTAT performs a correspondence analysis on a rectangular matrix. The singular value decomposition is used to compute row and column coordinates in a single configuration. These coordinates are popularly represented as a set of vectors for the columns and a set of points for the rows.

Biplots (Gabriel, 1971) are a singular value decomposition of a general $n \times p$ matrix. MDPREF (Carroll, 1972) is the same model except that the vectors (column coordinates) are standardized to have equal length. This is because Carroll developed the procedure for representing preferences with the vector model based on *n* subjects' preferences for *p* objects.

Procrustes Rotations

Procrustes rotations involve matching a source configuration to a target. SYSTAT offers two types of rotations. The first is a **classical orthogonal procrustes rotation** (Schönemann, 1966). This produces a fit by rotating and transposing axes and is especially suited for principal components and factor analyses.

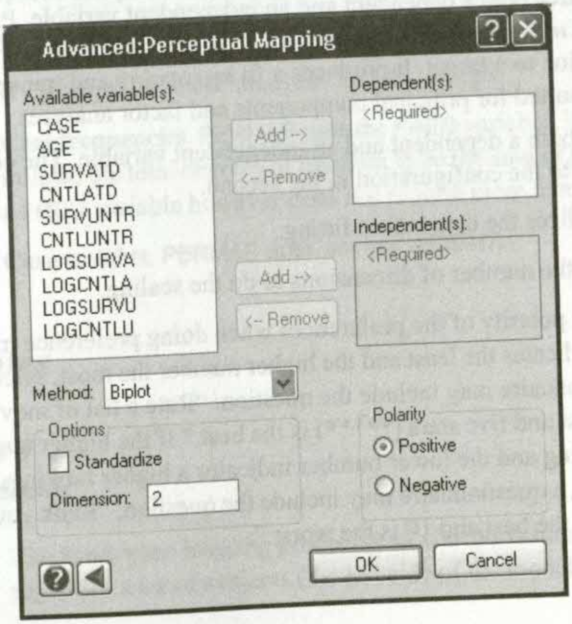
The second method, called **canonical rotation** in SYSTAT, is a general translation, rotation, reflection, and uniform dilation transformation that is ideally suited for multidimensional scaling and any procedure where location, scale, and orientation are arbitrary. This method is documented in Borg and Groenen (1997).

Perceptual Mapping in SYSTAT

Perceptual Mapping Dialog Box

To open the Perceptual Mapping dialog box, from the menus choose:

Advanced
  Perceptual Mapping...



Dependent(s). The dependent variables should be continuous or categorical numeric variables (for example, *income*).

Independent(s). The independent variables should be continuous or categorical numeric variables (grouping variables).

Method. The following methods are available:

- **Biplot.** Requires only a dependent variable. Biplots are a singular value decomposition of a general matrix.
- **Canonical.** Requires both a dependent and an independent variable. It relates an n-dimensional configuration to another. Canonical rotation is a general translation, rotation, reflection, and uniform dilation transformation that is ideally suited for multidimensional scaling and any procedure where location, scale, and orientation are arbitrary.
- **Circle.** Requires both dependent and independent variable(s). The columns of the first set are fit to the configuration in the second.
- **Ellipse.** Requires both dependent and independent variable(s). The columns of the first set are fit to the configuration in the second.
- **MDPREF.** Requires only a dependent variable. MDPREF is a biplot except that the vectors (column coordinates) are of the same unit length.
- **Procrustes.** Requires both a dependent and an independent variable. Procrustes rotation relates an n-dimensional configuration to another and involves matching a source configuration to a target. It produces a fit by rotating and transposing axes and is especially suited for principal components and factor analyses.
- **Vector.** Requires both a dependent and an independent variable. The columns of the first set are fit to the configuration in the second.

Standardize. Standardizes the data before fitting.

Dimension. Specifies the number of dimensions to do the scaling.

Polarity. Specifies the polarity of the preferences when doing preference mapping. If the smaller number indicates the least and the higher number the most, select Positive. For example, a questionnaire may include the question: "Rate a list of movies where one star (*) is the worst and five stars (*****) is the best." If the higher number indicates a lower ranking and the lower number indicates a higher ranking, select Negative. For example, a questionnaire may include the question, "Rank your favorite sports team where 1 is the best and 10 is the worst."

Using Commands

After selecting a file with USE filename, continue with:

```
PERMAP
     MODEL varlist or depvarlist = indvarlist
     ESTIMATE / METHOD = BIPLOT
                         MDPREF
                         VECTOR
                         CIRCLE
                         ELLIPSE
                         PROCRUSTES
                         CANONICAL,
                   STANDARDIZE,
                   DIMENSION = n,
                   POLARITY = POSITIVE
NEGATIVE
```

Usage Considerations

Types of data. PERMAP uses only rectangular data.

Print options. The output is standard for all PLENGTH options.

Quick Graphs. PERMAP produces Quick Graphs for every analysis.

Saving files. PERMAP does not save coordinates.

BY groups. PERMAP analyzes data by groups.

Case frequencies. PERMAP uses the FREQ variable, if present, to duplicate cases. This inflates the total degrees of freedom to be the sum of the number of frequencies. Using a FREQ variable however does not require more memory.

Case weights. PERMAP does not use WEIGHT.

Examples

Example 1
Vector Model

The Preference Mapping procedure of Carroll is implemented through a model that regresses a set of subjects (the left side of the model equation) onto the coordinates of

a set of objects (the right side of the model equation). The file *SYMP* contains coordinates from a multidimensional scaling of disease symptoms from Wilkinson, Blank, and Gruber (1996). It also contains, for a selected set of diseases, indicators for the presence or absence of a symptom. These are informal ratings.

The input for fitting the vector model to the data is:

```
USE SYMP
IDVAR SYMPTOM$
PERMAP
          MODEL LYME MALARIA YELLOW RABIES FLU = DIM1 DIM2
          ESTIMATE / METHOD=VECTOR
```

The output is:

```
Configuration has been centered prior to fitting
External unfolding via vector model
```

Goodness of Fit for Subjects

Subject	R-square	F-ratio	df		p-value
1	0.007	0.056	2	15	0.946
2	0.029	0.226	2	15	0.800
3	0.173	1.574	2	15	0.240
4	0.059	0.474	2	15	0.632
5	0.079	0.642	2	15	0.540

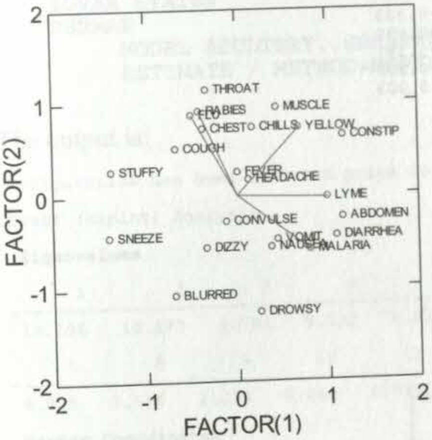
Regression Coefficients for Subjects

	1	2	3
1	0.000	0.057	-0.002
2	0.000	0.096	-0.070
3	0.000	0.170	0.177
4	0.000	-0.073	0.161
5	0.000	-0.084	0.147

Subject Coordinates

	1	2
1	0.999	-0.032
2	0.806	-0.592
3	0.692	0.722
4	-0.415	0.910
5	-0.496	0.869

External Unfolding Vector Model



Example 2
Circle Model

The circle model places the diseases near the symptoms they most involve.

The input is:

```
USE SYMP
IDVAR SYMPTOM$
PERMAP
MODEL LYME MALARIA YELLOW RABIES FLU = DIM1 DIM2
ESTIMATE / METHOD=CIRCLE
```

The output is:

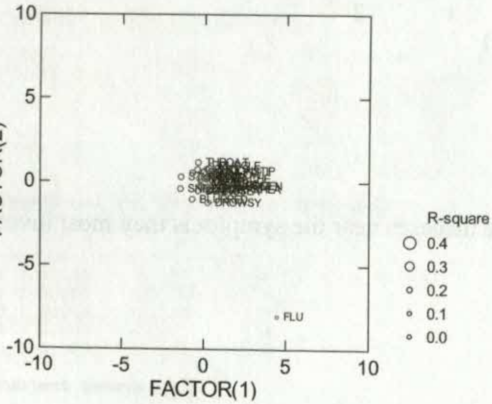
```
Configuration has been centered prior to fitting
External unfolding via circular ideal point model
```

Goodness of Fit for Subjects

Subject	R-square	F-ratio	df		p-value	
1	0.271	1.735	3	14	0.206	
2	0.385	2.926	3	14	0.071	
3	0.265	1.685	3	14	0.216	
4	0.257	1.615	3	14	0.231	
5	0.079	0.401	3	14	0.755	Anti-Ideal

```
Regression Coefficients for Subjects

   |       1        2        3        4
---+-----------------------------------
 1 |   0.379    0.001   -0.046   -0.380
 2 |   0.449    0.029   -0.123   -0.450
 3 |   0.191    0.142    0.155   -0.191
 4 |   0.334   -0.123    0.122   -0.335
 5 |  -0.009   -0.083    0.148    0.009

Subject Coordinates

   |       1        2
---+----------------
 1 |   0.001   -0.061
 2 |   0.033   -0.136
 3 |   0.370    0.405
 4 |  -0.183    0.182
 5 |   4.475   -8.030
```

External Unfolding Circular Model



Example 3
Internal Model

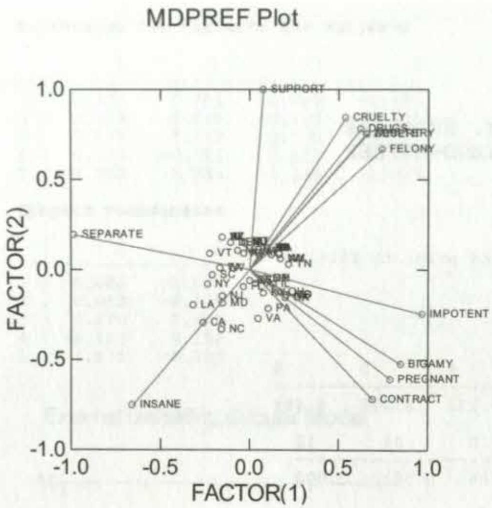
The *DIVORCE* file includes grounds for divorce in the United States in 1971. It is adapted from Wilkinson, Blank, and Gruber (1996), and is originally from Long (1971). We will do an MDPREF analysis on these data to plot the rows and columns in a common space.

The input is:

```
USE DIVORCE
IDVAR STATE$
PERMAP
     MODEL ADULTERY..SEPARATE
     ESTIMATE / METHOD=MDPREF
```

The output is:

```
Configuration has been centered prior to fitting

MDPREF (Biplot) Analysis

Eigenvalues

      1        2        3        4        5        6
-------------------------------------------------------
 19.196   18.277    9.792    9.132    8.035    6.483

      7        8        9       10       11       12
-------------------------------------------------------
  4.825    3.595    2.278    0.866    0.581    0.000

Vector Coordinates

    |      1        2
----+------------------
  1 |  0.657    0.754
  2 |  0.540    0.842
  3 |  0.661    0.751
  4 |  0.078    0.997
  5 |  0.746    0.666
  6 |  0.969   -0.248
  7 |  0.793   -0.610
  8 |  0.626    0.780
  9 |  0.694   -0.720
 10 | -0.657   -0.754
 11 |  0.851   -0.525
 12 | -0.982    0.191

Object Coordinates

    |      1        2
----+------------------
  1 |  0.102    0.117
  2 |  0.048    0.092
  3 | -0.032    0.088
  4 | -0.152    0.178
  . |    .        .
  . |    .        .
 49 | -0.165    0.007
 50 |  0.176    0.057
```



The biplot looks similar, with all of the grounds for divorce vectors approximately equal in length, because the original data have comparable variances on these variables.

Example 4
Procrustes Rotation

In a profound but seldom-cited dissertation, Wilkinson (1975) scaled perceptions of cars and dogs among car club and dog club members. The file *CARDOG* contains the INDSCAL (Individual Differences Scaling) configurations of the scalings of cars and dogs. Wilkinson paired cars and dogs by using the subject's responses on additional rating scales of attributes. INDSCAL dimensions, on the other hand, are claimed to have an intrinsic canonical orientation that ordinarily precludes rotation (see the references in "Multidimensional Scaling"). The question here, then, is whether a procrustes rotation guided by the extrinsically based pairings will change the original INDSCAL configurations. We will rotate cars to dogs.

The input is:

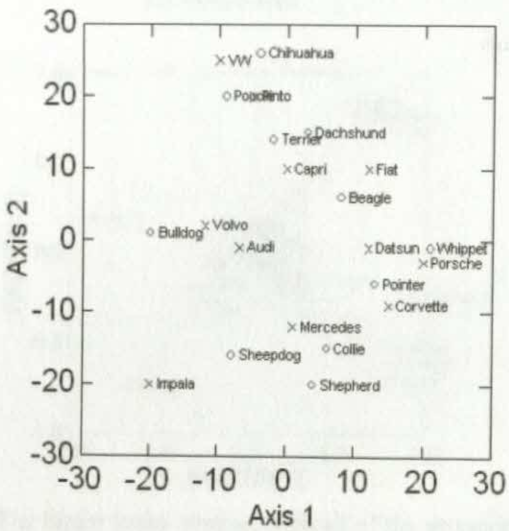
```
USE CARDOG
PERMAP
     MODEL C1,C2 = D1,D2
     ESTIMATE/METHOD=PROCRUSTES
```

The output is:

```
Orthogonal Procrustes Rotation

 Rotation Matrix T

   |      1        2
---+-----------------
 1 |  0.984    0.177
 2 | -0.177    0.984

 Target (X) Coordinates

    |      1         2
----+------------------
  1 |  21.000    -1.000
  2 |  13.000    -6.000
  3 |  -4.000    26.000
  4 |  -9.000    20.000
  5 |   3.000    15.000
  6 |  -2.000    14.000
  7 | -20.000     1.000
  8 |  -8.000   -16.000
  9 |   4.000   -20.000
 10 |   6.000   -15.000
 11 |   8.000     6.000

 Rotated (Y) Coordinates

    |      1         2
----+------------------
  1 |  20.216     0.577
  2 |  16.353    -6.211
  3 | -14.256    22.842
  4 |  -8.452    18.803
  5 |  10.047    11.961
  6 |  -1.765     9.843
  7 | -12.165    -0.149
  8 | -16.156   -23.216
  9 |   3.102   -11.635
 10 |  -6.714    -2.220
 11 |  11.988     1.134
```

The rotation matrix in the output is nearly an identity matrix. Unlike the nonmetric multidimensional scalings in Wilkinson's dissertation, which required rotation to a common orientation, the INDSCAL analyses recovered the apparently canonical dimensions. These were agile-clumsy (horizontal) and big-small (vertical).

In place of the procrustes output, which normally consists of separate scatterplots of the two sets, we present a plot of the superimposed configurations.



Computation

Algorithms

The algorithms are documented in the Statistical Background section above. Most involve a singular value decomposition computed in the standard manner.

Missing data

Cases and variables with missing data are omitted from the calculations.

References

Bennet, J. F. and Hays, W. L. (1960). Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika*, 25, 27–43.

Borg, I. and Groenen, P. J. F. (1997). *Modern multidimensional scaling: Theory and applications.* New York: Springer Verlag.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, S. B. Nerlove (eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences, Vol. 1*, 105–155. New York: Seminar Press.

Cliff, N. (1973). Scaling. *Annual Review of Psychology*, 24, 473–506.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review,* 57, 145–158.

Dow, B. M. (1990). Nested maps in macaque monkey visual cortex. In K. N. Leibovic (ed.), *Science of vision*, 84–124. New York: Springer Verlag.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika,* 58, 453–467.

Green, P. E. and Tull, D. S. (1975). *Research for marketing decisions,* 3rd ed. Englewood Cliffs, N. J.: Prentice Hall.

Kosslyn, S. M. (1981). *Image and mind*. Cambridge: Harvard University Press.

Lilien, G. L., Kotler, P., and Moorthy, K. S. (1992). *Marketing models*. Englewood Cliffs, N.J.: Prentice Hall.

Long, L. H. (ed.) (1971). *The world almanac*. New York: Doubleday.

Milgram, S. and Jodelet, D. (1976). Psychological maps of Paris. In H. M. Proshansky, W. H. Itelson, and L. G. Revlin (eds.), *Environmental Psychology*. New York: Holt, Rinehart, and Winston.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana, Ill.: University of Illinois Press.

Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika,* 31, 1–10.

* Schwartz, E. J. (1981). Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Visual Research,* 20, 645–669.

Shepard, R. N. and Cooper, L. A. (1982). *Mental images and their transformations*. MA: MIT Press, Cambridge.

Wilkinson, L. (1975). The effect of involvement on similarity and preference structures. Unpublished Ph.D. dissertation, Yale University.

Wilkinson, L., Blank, G., and Gruber, C. (1996). *Desktop data analysis with SYSTAT*. Upper Saddle River, N.J.: Prentice-Hall.

Wish, M., Deutsch, M., and Biener, L. (1972). Differences in perceived similarity of nations. In R. N. Shepard, A. K. Romney, S. B. Nerlove (eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences, Vol. 2*, 289–313. New York: Seminar Press.

(* indicates additional reference.)

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, S. B. Nerlove (eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences. Vol. I*, 105–155. New York: Seminar Press.
Cliff, N. (1973). Scaling. *Annual Review of Psychology*, 24, 473–506.
Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, 145–158.
Dow, B. M. (1990). Nested maps in macaque monkey visual cortex. In K. N. Leibovic (ed.), *Science of Vision*, 84–124. New York: Springer Verlag.
Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 453–467.
Green, P. E. and Tull, D. S. (1970). *Research for marketing decisions* 3rd ed. Englewood Cliffs, N. J.: Prentice Hall.
Kosslyn, S. M. (1980). *Image and mind*. Cambridge: Harvard University Press.
Lilien, G. L., Kotler, P., and Moorthy, K. S. (1992). *Marketing models*. Englewood Cliffs, N.J.: Prentice Hall.
Lon, L. H. (ed.) (1971). *The world almanac*. New York: Doubleday.
Milgram, S. and Jodelet, D. (1976). Psychological maps of Paris. In H. M. Proshansky, W. H. Ittelson, and L. G. Rivlin (eds.), *Environmental Psychology*. New York: Holt, Rinehart, and Winston.
Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, Ill.: University of Illinois Press.
Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31, 1–10.
Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research*, 20, 645–669.
Shepard, R. N. and Cooper, L. A. (1982). *Mental images and their transformations*. MA: MIT Press, Cambridge.
Wilkinson, L. (1975). *The effect of involvement on similarity and preference structures*. Unpublished Ph.D. dissertation, Yale University.
Wilkinson, L., Blank, G., and Gruber, C. (1996). *Desktop data analysis with SYSTAT*. Upper Saddle River, N.J.: Prentice-Hall.
Wish, M., Deutsch, M., and Biener, L. (1972). Differences in perceived similarity of nations. In R. N. Shepard, A. K. Romney, S. B. Nerlove (eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences*. New York: Seminar Press.

(indicates additional reference)

Chapter
2

Power Analysis

Gerard E. Dallal, Michael Pechnyo, and Rick Marcantonio
(revised by Naresh A. Raj)

For a specific experimental design, power analysis (POWER) explores the relationship between sample size and the probability of achieving statistical significance. Available experimental designs include:

- comparing a single proportion to a value
- equality of two independent proportions
- comparing a correlation coefficient to a value
- equality of two correlation coefficients
- z-tests (one-sample and two-sample)
- t-tests (one-sample, paired, and two-sample)
- one-way ANOVA
- two-way ANOVA

Power calculations for other designs can be performed using generic power analysis. In this case, specify the degrees of freedom and the non-centrality parameter to perform the analysis. This approach can be used for general factorial designs, randomized block designs, and fixed effect regression, as well as many other designs.

In general, power depends on the parameters of the population(s) involved, the probability of making an error, and the size of the sample(s). For a fixed error rate and set of population parameters, you can either find the sample size needed to achieve a specific power level or find the power corresponding to a specific sample size. You can also find the power for each sample in a range of sample sizes.

The results can be saved to a data file for further analysis. For a fixed sample size or power level, you can vary the effect size or alpha level and append the saved

estimates together in a single file. The resulting file can be used to create custom plots depicting relationships between power, alpha, effect size, and sample size using power curve overlays, contour plots, mosaic plots, or power surfaces.

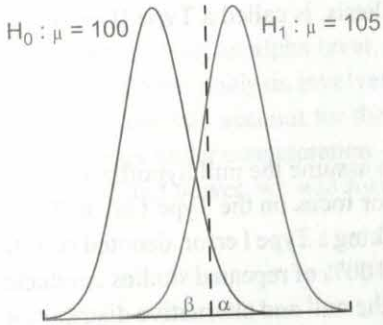
Statistical Background

In inferential statistics, a researcher collects data from a sample to test hypotheses about the population from which the sample is drawn. For example, you could compare the mean values for a continuous variable measured for each of two samples to determine if the (unobserved) populations represented by the samples differ.

Hypothesis tests typically involve two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis, usually denoted H_0, asserts that the treatment has no effect. In other words, the parameter for the population in question does not differ from the parameter for the general population. In hypothesis testing, we evaluate the likelihood of obtaining the observed sample estimate of the parameter assuming the null hypothesis is true. If the probability of obtaining the observed statistic indicates a very rare occurrence, we reject the null hypothesis in favor of the other hypothesis, known as the alternative hypothesis. The alternative hypothesis, denoted H_1, may suggest either a specific value or a range of values for the parameter.

The likelihood of observing a particular value given that the null hypothesis is true is called the ***p-value*** of a test. The *p-value* corresponds to a probability calculated using a distribution associated with the hypothesis test. "Small" *p-values* lead to rejection of the null hypothesis in favor of the alternative. However, the definition of "small" varies from researcher to researcher.

The distribution used to determine the *p-value* varies from test to test, but the concepts underlying power analysis generalize across tests. To simplify our discussion, we will focus on the one-sample t-test. Consider the two distributions shown below.



The distribution on the left suggests that the mean equals 100. This corresponds to a null hypothesis of μ=100. We will refer to this distribution as the null distribution. In contrast, the other distribution is shifted to the right, resulting in a center of 105. We will refer to this distribution as the alternative distribution. Both distributions have a standard deviation of 10; they vary only in their center.

Error Types

Hypothesis testing assumes the null hypothesis is true and sets out to find evidence to the contrary. In fact, the null hypothesis could be true or false. Moreover, the researcher could decide to reject the null hypothesis in favor of the alternative hypothesis or retain the null hypothesis due to a lack of evidence against it. The state of the null hypothesis and the decision of the researcher form a two-by-two table of possible outcomes of the hypothesis test.

		Researcher's Conclusion	
		Retain Ho	**Reject Ho**
Actual State	**Ho true**	correct decision	Type I error
	Ho false	Type II error	correct decision

Consider the case where the null hypothesis is actually true. If the researcher retains the null hypothesis, the decision is correct. Alternatively, a decision to reject the null results in an error, commonly referred to as a Type I error.

Now shift attention to the case in which the null hypothesis is actually false. The researcher still faces the same decisions. However, rejecting the null hypothesis would

be a correct action in this case. An error occurs when the null hypothesis is not rejected. This error, retaining a false null hypothesis, is called a Type II error.

Probabilities of Errors

When conducting a hypothesis test, we assume the null hypothesis is true, so attempts to control the chance of making an error focus on the Type I error. The researcher defines a probabilistic tolerance for making a Type I error, denoted α. If the null is true, we expect to make a Type I error in $\alpha*100\%$ of repeated studies conducted in the same fashion. In the previous illustration of the null and alternative distributions for the one-sample t-test, the dashed vertical line defines an area under the null distribution equal to α. Because we either reject or retain the null hypothesis, the probability of retaining the null equals $1-\alpha$.

The probability of obtaining the observed data given that the null is true, the *p-value*, is compared to the probability of making a Type I error. If the *p-value* is smaller than the alpha level, we reject the null in favor of the alternative.

However, the null hypothesis may actually be false. Under these circumstances, the researcher may make a Type II error. The probability of making this error equals β and corresponds to the area to the left of the vertical line in our one-sample t-test. The probability of a correct decision is $1-\beta$.

We would like the probability of a correct decision to be very high. This suggests that we want α and β to be as small as possible. However, an inverse relationship exists between α and β; as α gets smaller, β gets bigger. This results in a tradeoff between the error tolerances. If you want to make the probability of a Type I error very small, you increase the probability of a Type II error.

Power

The **Power** of a test equals the probability of making a correct decision when the null hypothesis is false. This value corresponds to $1-\beta$. Look at the two distributions displayed previously. The area under the alternative curve to the right of the critical value corresponds to power.

How can we maximize power? The following four general components affect the power of a statistical test:

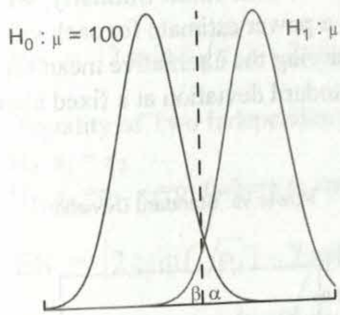
- **Effect size.** Characteristics of the population under consideration.
- **Alpha and Beta.** Tolerance for errors in decisions.

- **Sample Size.** A characteristic of the sample used in the study.

Given an effect size, an alpha level, and a sample size, power is completely determined. Power analysis involves finding an optimal combination of these components that can account for the available resources for a study while addressing the hypotheses under consideration. To explore the relationship between each of these components and power, we will focus on each in turn.

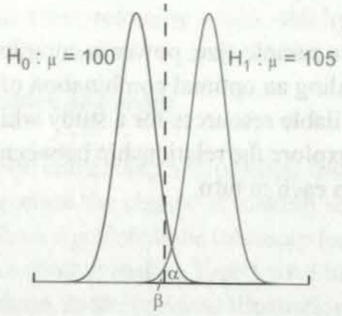
Effect Size

Return to the one-sample t-test introduced earlier. Suppose that the alternative distribution is centered at 107 instead of 105. Holding all other characteristics constant, the alternative distribution gets shifted to the right.



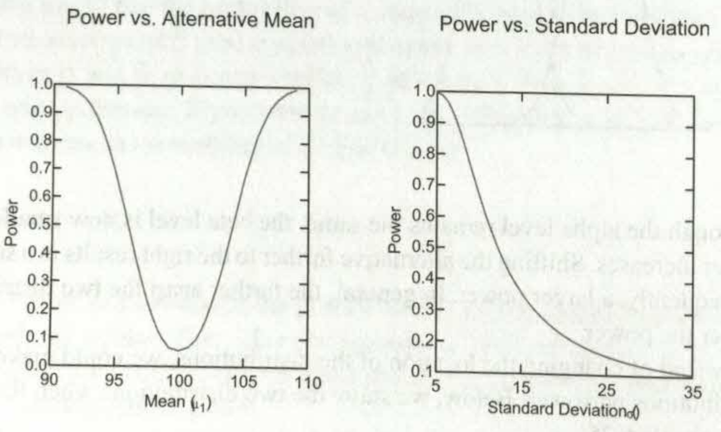
Although the alpha level remains the same, the beta level is now smaller. As a result, power increases. Shifting the alternative further to the right results in a smaller beta and consequently, a larger power. In general, the farther apart the two distributions, the higher the power.

Instead of changing the location of the distributions, we could make the distributions narrower. Below, we show the two distributions when the variance of each equals 6.25.



Beta is now very small, resulting in a large power. As variance decreases, power increases.

The previous examples illustrated the situation for one mean and one variance. We could find the power corresponding to each example. Instead, let the alternative mean vary over a range and calculate the power for each value. Similarly, we can vary the standard deviation over a range to derive a power estimate for each value. Below, we depict the power curves resulting from varying the alternative mean for a fixed standard deviation and for varying the standard deviation at a fixed alternative mean.



In the left plot, as the alternative mean moves farther from the null mean (100), power increases. In the right plot, as the standard deviation increases, power decreases.

Whether changing the distance between the means or changing the variance of the distributions, as the proportion of overlap between the two distributions decreases, power increases. We need a measure that captures this relationship.

Effect size (ES) is commonly used as a measure of the difference between the null and alternative distributions. When the null hypothesis is true, the effect size equals 0. When the null is false, the effect size assumes a nonzero value. The larger the effect size, the larger the power. The larger the effect size, the smaller the sample needed to find it.

The computation of effect size depends on hypothesized values for parameters involved in the statistical test. Because each test relies on different parameters, no single formula for effect size exists. Instead, each test has a corresponding effect size computation.

- Single Proportion

  H_0: $\pi = \pi_0$
  H_1: $\pi = \pi_1$ (where $\pi_0 \neq \pi_1$)

$$ES = \left|2\,\mathrm{asin}(\sqrt{\pi_1}) - 2\,\mathrm{asin}(\sqrt{\pi_0})\right|$$

- Equality of Two Independent Proportions

  H_0: $\pi_1 = \pi_2$
  H_1: $\pi_1 = p_1$; $\pi_2 = p_2$ (where $p_1 \neq p_2$)

$$ES = \left|2\,\mathrm{asin}(\sqrt{p_1}) - 2\,\mathrm{asin}(\sqrt{p_2})\right|$$

- Single Correlation Coefficient

  H_0: $\rho = 0$
  H_1: $\rho = \rho_1$ (where $\rho_1 \neq 0$)

$$ES = \rho_1$$

  H_0: $\rho = \rho_0$
  H_1: $\rho = \rho_1$ (where $\rho_0 \neq \rho_1$)

$$ES = \frac{\left|\ln\left(\frac{1+\rho_1}{1-\rho_1}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right|}{2}$$

- Equality of Two Correlations

  H_0: $\rho_1 = \rho_2$
  H_1: $\rho_1 = c_1$; $\rho_2 = c_2$ (where $c_1 \neq c_2$)

$$ES = \frac{\left|\ln\left(\frac{1+c_1}{1-c_1}\right) - \ln\left(\frac{1+c_2}{1-c_2}\right)\right|}{2}$$

- One-sample t-test
  H_0: $\mu = \mu_0$
  H_1: $\mu = \mu_1$ (where $\mu_0 \neq \mu_1$)

$$ES = \frac{|\mu_1 - \mu_0|}{\sigma}$$

- Paired t-test
  H_0: $\mu_{diff} = 0$
  H_1: $\mu_{diff} = \Delta$ (where $\Delta \neq 0$)

$$ES = \frac{|\Delta|}{\sigma_{diff}}$$

- Two-sample t-test
  H_0: $\mu_1 = \mu_2$
  H_1: $\mu_1 = m_1$; $\mu_2 = m_2$ (where $m_1 \neq m_2$)

$$ES = \frac{|m_1 - m_2|}{\sigma}$$

- One-way ANOVA
  H_0: $\mu_1 = \mu_2 = ... = \mu_I$
  H_1: $\mu_1 = m_1$; $\mu_2 = m_2$;...$\mu_I = m_I$

$$ES = \frac{\sqrt{\frac{\sum_{i=1}^{I}(m_i - \mu)^2}{I}}}{\sigma}$$

- Two-way ANOVA

H_0: $\mu_1 = \mu_2 = ... = \mu_R$
H_1: $\mu_1 = m_1$; $\mu_2 = m_2$;...$\mu_R = m_R$

$$ES = \frac{\sqrt{\frac{\sum_{i=1}^{R} (m_i - \mu)^2}{R}}}{\sigma}$$

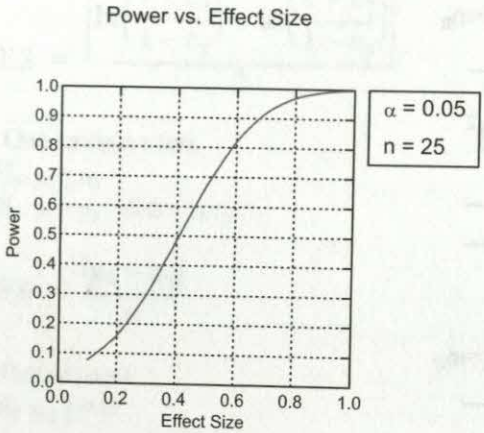
H_0: $\mu_1 = \mu_2 = ... = \mu_C$
H_1: $\mu_1 = m_1$; $\mu_2 = m_2$;...$\mu_C = m_C$

$$ES = \frac{\sqrt{\frac{\sum_{j=1}^{C} (m_j - \mu)^2}{C}}}{\sigma}$$

H_0: $\mu_{11} = \mu_{12} = ... = \mu_{RC}$
H_1: $\mu_{11} = m_{11}$; $\mu_{12} = m_{12}$;...$\mu_{RC} = m_{RC}$

$$ES = \frac{\sqrt{\frac{\sum_{i,j}^{R,C} (m_{ij} - \mu)^2}{RC}}}{\sigma}$$

Instead of plotting power as a function of the mean difference or as a function of the variance, we can combine the two and plot power as a function of the effect size.



As the effect size increases, power increases. As the discrepancy between the alternative and the null increases, power increases.

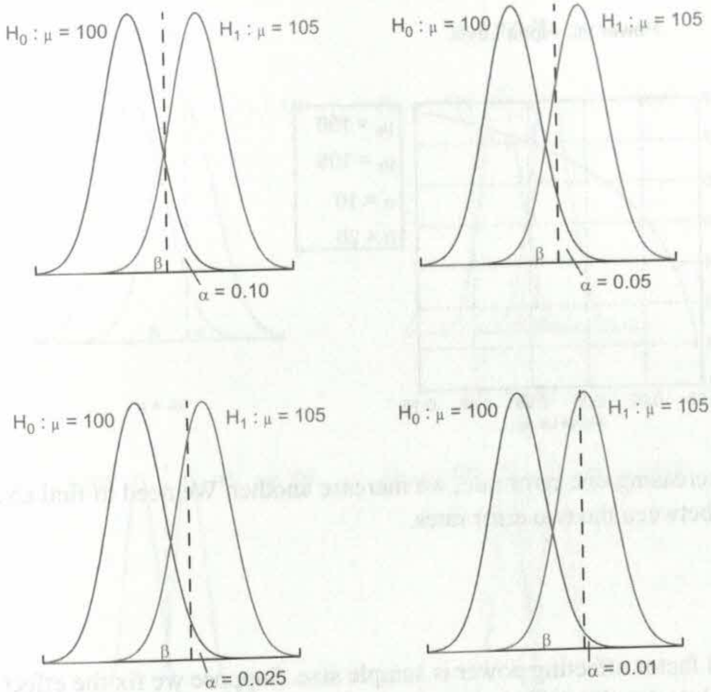
Cohen (1988) offers the following guidelines for classifying effect sizes as small, medium, or large based on the type of test under consideration.

Design	Effect Size		
	Small	**Medium**	**Large**
Proportions	0.20	0.50	0.80
Correlation Coefficients	0.10	0.30	0.50
t-tests	0.20	0.50	0.80
ANOVA	0.10	0.25	0.40

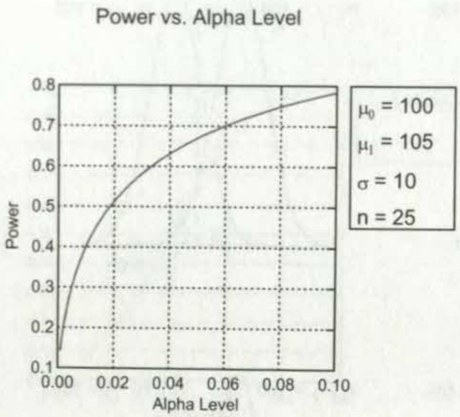
However, as with any rule of thumb, these values should be interpreted cautiously. The effect size depends on the study under consideration, and should not be set by mere conventions.

Alpha

Alpha represents an error rate, and as such, should be set as small as possible. However, alpha and beta are intimately related. Consider the previous example, but allow alpha to change instead of the effect size.



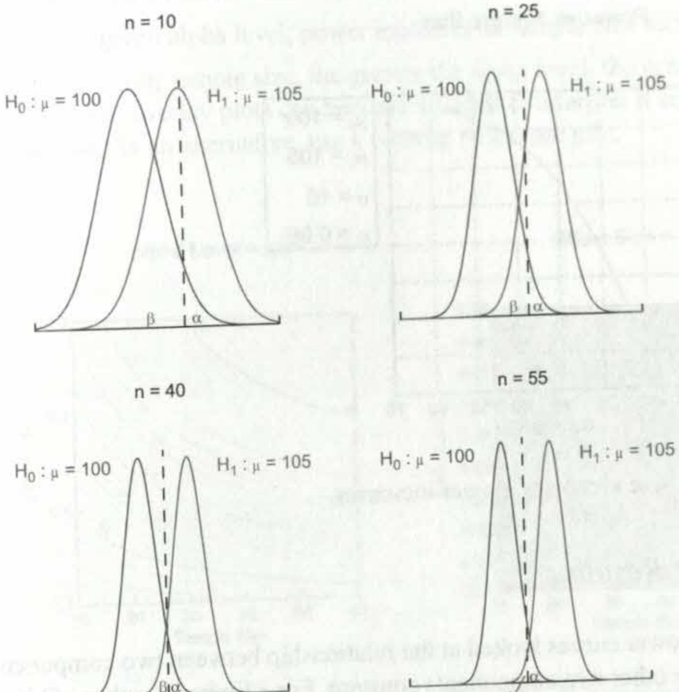
As alpha gets smaller, beta gets bigger. We can compute the power for several alpha levels and plot the results:



So by decreasing one error rate, we increase another. We need to find an acceptable balance between the two error rates.

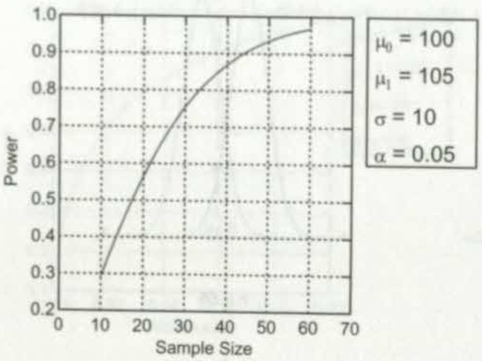
Sample Size

The final factor affecting power is sample size. Suppose we fix the effect size at 0.5 and the alpha level at 0.05. Let's return to our two distributions and look at the curves for four different sample sizes.



As the sample size increases, the distributions become more peaked. The distributions do not change in location, but do change in spread. The standard deviation of these distributions equals the standard error of the distribution of sample means, which depends on the sample size. For a fixed effect size and alpha level, we can compute the power for a range of sample sizes.

Power vs. Sample Size

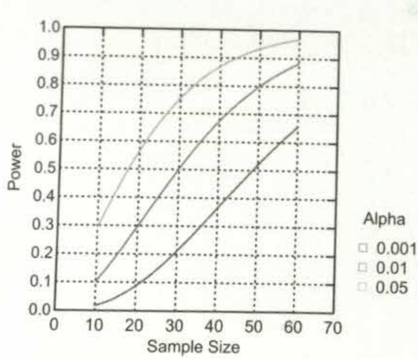


As sample size increases, power increases.

Displaying Power Results

Previous power curves looked at the relationship between two components while holding the other two components constant. For a limited number of alternatives, overlaying power curves allows you to examine the relationship between three components holding only the fourth constant.

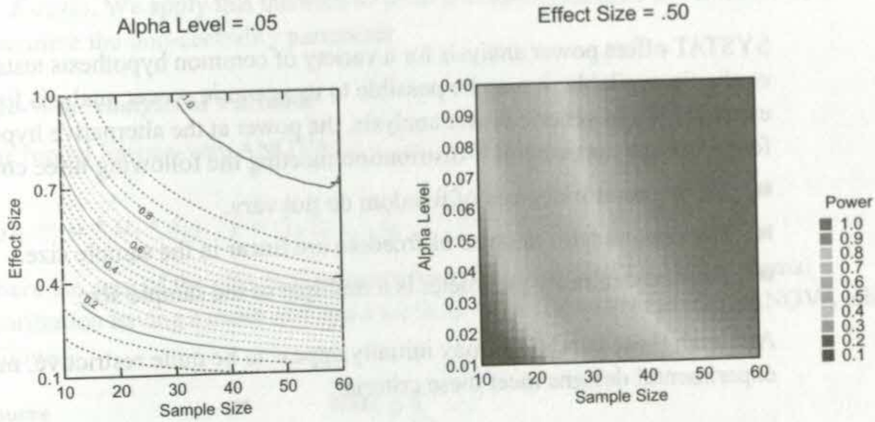
Power Curves (Effect Size = .50)



This plot demonstrates:

- for any given alpha level, power increases as sample size increases.
- for any given sample size, the greater the alpha level, the greater the power.

However, overlay plots can become difficult to interpret if several curves are desired. As an alternative, use a contour or mosaic plot.



The contour plot on the left looks at the relationship between power, effect size, and sample size for a fixed alpha level (0.05). The contour lines represent power. Close contours indicate rapid increases in power; distant contours indicate gradual increases. This plot can be used to find:

- the power for an effect size/sample size pair.
- the effect size for a power/sample size pair.
- the sample size for a power/effect size pair.

For example, at an effect size of 0.70, to obtain a power of 0.80, we would need approximately 20 observations. However, to detect a smaller difference between the null and alternative hypotheses, such as that described by an effect size of 0.40, we need about 50 observations.

The mosaic plot on the right offers an alternative representation of the relationship between three of the four components. In this plot, color gradients represent power. In contrast to the contour plot, we fix the effect size instead of the alpha level and allow the power, the sample size, and the alpha level to vary.

Instead of collapsing three factors into a two-dimensional plot, you can display the relationship using a three-dimensional surface. Using the Dynamic Explorer, you can rotate this graph to view the surface from any desired perspective, allowing you to find areas of rapid increases in power. See the examples for an illustration of a power surface.

Generic Power Analysis

SYSTAT offers power analysis for a variety of common hypothesis tests. For tests not explicitly available, it may be possible to use generic power analysis for planning of experiments. In generic power analysis, the power at the alternative hypothesis can be found using a non-central F-distribution meeting the following three conditions:

- The numerator degrees of freedom do not vary.
- The denominator degrees of freedom are linear in the sample size.
- The non-centrality parameter is a multiple of the sample size.

Although these conditions may initially appear to be quite restrictive, many experimental designs meet these criteria.

Non-centrality Parameters

The statistical literature contains many definitions of the non-centrality parameter of the non-central F-distribution. The differences typically involve a square root, a factor of (numerator degrees of freedom +1), or a factor of 2. We follow the notation of Kendall, Ord, Stuart and Arnold (1999):

- The sum of the squares of d independent normal variables with arbitrary means and unit variances is said to follow a non-central chi-square distribution with d degrees of freedom and non-centrality parameter equal to the sum of the squared means.
- The ratio of a non-central chi-square variable with d_1 degrees of freedom and non-centrality parameter lambda, divided by d_1, to an independent central chi-square variable with d_2 degrees of freedom, divided by d_2, is said to follow a non-central F-distribution with d_1 numerator degrees of freedom, d_2 denominator degrees of freedom, and non-centrality parameter lambda.

Tang (1938) uses one-half of lambda as the non-centrality parameter. Scheffé (1959) defines his value to be the square root of lambda. Odeh and Fox (1991) divide lambda by the numerator df plus 1 before taking the square root.

Graybill (1961, Theorem 11.16) notes that a non-centrality parameter can be obtained as the numerator degrees of freedom times (the difference between the numerator expected mean square and the error variance) divided by the error variance, where the error variance is given by the expected mean square of the denominator of the *F-ratio*. We apply this theorem to several common designs to demonstrate how to determine the non-centrality parameter.

One-Way Analysis of Variance

The model for one-way ANOVA is:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, n$$

where the sum of the effects (α_i) equals 0 and the error terms follow a normal distribution having a mean of 0 and a variance of σ^2. The corresponding ANOVA table follows:

Source	df	EMS
Between Groups	I-1	$\sigma^2 + \frac{n\sum \alpha_i^2}{I-1}$
Error (Within Groups)	nI-I	σ^2

Using the expected mean squares, the non-centrality parameter equals:

$$\lambda = \frac{(I-1)\frac{n\sum \alpha_i^2}{I-1}}{\sigma^2} = n\frac{\sum \alpha_i^2}{\sigma^2}$$

Notice that the numerator degrees of freedom are constant and the denominator degrees of freedom have a linear relationship with the sample size. As a result, one-way analysis of variance could be analyzed using the generic method.

Two-Way Analysis of Variance

The model for two-way ANOVA is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},\ i = 1, 2, \ldots, I, j = 1, 2, \ldots, J, k = 1, 2, \ldots, n$$

where the sum of each set of effects equals 0 and the error terms follow a normal distribution having a mean of 0 and a variance of σ^2. The corresponding ANOVA table follows:

Source	**df**	**EMS**
Factor A	I-1	$\sigma^2 + \frac{nJ\sum \alpha_i^2}{I-1}$
Factor B	J-1	$\sigma^2 + \frac{nI\sum \beta_j^2}{J-1}$
Interaction	(I-1)(J-1)	$\sigma^2 + \frac{n\sum (\alpha\beta)_{ij}^2}{(I-1)(J-1)}$
Error (Within Groups)	IJ(n-1)	σ^2

The non-centrality parameter for factor A equals:

$$\lambda_A = \frac{(I-1)\frac{nJ\sum \alpha_i^2}{I-1}}{\sigma^2} = nJ\frac{\sum \alpha_i^2}{\sigma^2}$$

For factor B, the non-centrality parameter is:

$$\lambda_B = \frac{(J-1)\frac{nI\sum \beta_j^2}{J-1}}{\sigma^2} = nI\frac{\sum \beta_j^2}{\sigma^2}$$

The non-centrality parameter for the interaction between factors A and B equals:

$$\lambda_{A\times B} = \frac{(I-1)(J-1)\dfrac{n\sum(\alpha\beta)_{ij}^2}{(I-1)(J-1)}}{\sigma^2} = n\frac{\sum(\alpha\beta)_{ij}^2}{\sigma^2}$$

For three and higher way ANOVA models, the computations are similar.

Randomized Blocks Designs

The model is:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, n$$

where the sum of the treatment effects equals 0, the sum of the block effects equals 0, and the error terms follow a normal distribution having a mean of 0 and a variance of σ^2. The corresponding ANOVA table follows:

Source	**df**	**EMS**
Treatments	I-1	$\sigma^2 + \frac{n\sum\tau_i^2}{(I-1)}$
Blocks	n-1	$\sigma^2 + \frac{I\sum\beta_j^2}{n-1}$
Error	(I-1)(n-1)	σ^2

The non-centrality parameter for treatment effects equals:

$$\lambda = \frac{(I-1)\dfrac{n\sum\tau_i^2}{I-1}}{\sigma^2} = n\frac{\sum\tau_i^2}{\sigma^2}$$

Simple Linear Regression

The linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \ldots, N$$

where the error terms follow a normal distribution having a mean of 0 and a variance of σ^2. The corresponding ANOVA table is:

Source	**df**	**EMS**
Regression	1	$\sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$
Error	N-2	σ^2

The non-centrality parameter for regression effects equals:

$$\lambda = \frac{\beta_1^2 \sum_{i=1}^{N} (x_i - \bar{x})^2}{\sigma^2}$$

If we center the *x* values at zero, the non-centrality parameter becomes

$$\lambda = \frac{\beta_1^2 \sum_{i=1}^{N} x_i^{\,2}}{\sigma^2}$$

Consider the situation in which *p* possible values of *x* are possible. If we take *n* observations at each value of *x*, the non-centrality parameter becomes:

$$\lambda = \frac{\beta_1^2 \sum_{i=1}^{np} x_i^{\,2}}{\sigma^2} = \frac{\beta_1^2 \sum_{i=1}^{p} n(x_i^{\,2})}{\sigma^2} = n\frac{\beta_1^2 \sum_{i=1}^{p} x_i^{\,2}}{\sigma^2}$$

Consequently, generic power analysis can also be used for simple linear regression.

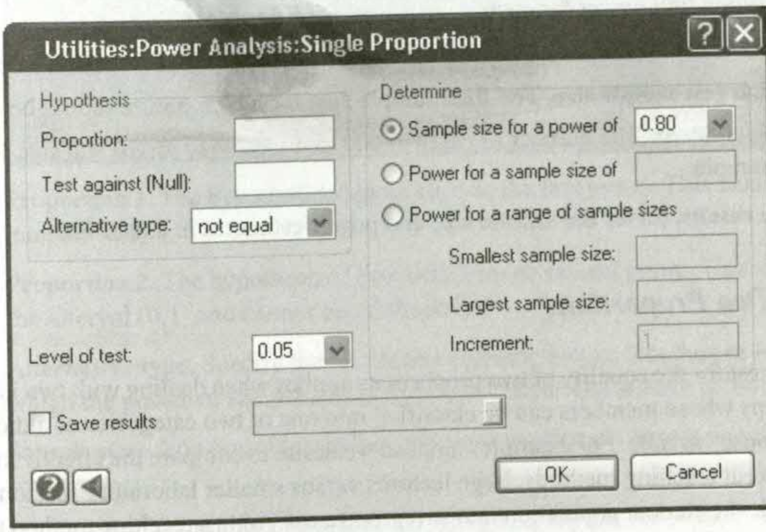
Power Analysis in SYSTAT

Single Proportion

Use the test for a single proportion for a situation involving one group of subjects whose members can be classified into one of two categories of a dichotomous response variable, such as successes and failures. For instance, in a public opinion poll, we could ask people if they approve or disapprove of the current political administration. If sentiment was evenly split, 0.50 of the respondents should respond in each category. However, we hypothesize that recent events will sway opinions to be more favorable, leading to a 0.60 approval rating.

To open the Single Proportion dialog box, from the menus choose:

Utilities
  Power Analysis
    Single Proportion...



Proportion. Enter the hypothesized value of the proportion according to the alternative hypothesis. This value must lie in the interval (0,1).

Test against (Null). Enter the hypothesized value of the proportion according to the null hypothesis. This value must lie in the interval (0,1) and differ from the value for Proportion.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Level of test. Specify the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the proportion for each hypothesis determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each:

- **Smallest sample size.** The initial sample size for which power should be computed.
- **Largest sample size.** The final sample size for which power should be computed.
- **Increment.** The number of cases to add to a sample to arrive at a size for the next sample.

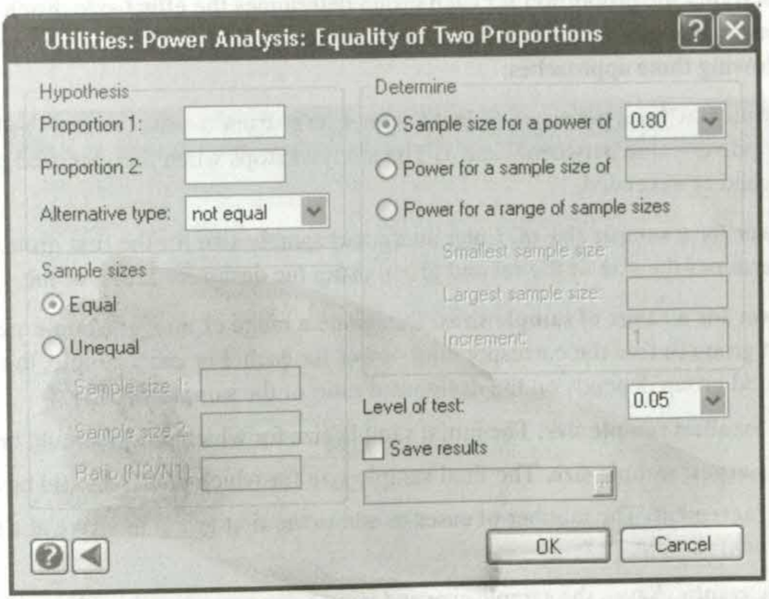
Save results. Saves the sample size and power estimates in a data file.

Equality of Two Proportions

The test for the equality of two proportions applies when dealing with two independent groups whose members can be classified into one of two categories of a dichotomous response variable. For example, suppose we desire to compare the effectiveness of two different teaching methods, large lectures versus smaller laboratory sessions. We will divide the student population into two groups, assigning a teaching method to each. At the end of the semester, we will record the number of students passing and the number failing using a common exam. The null hypothesis asserts that the proportion passing will be the same in the two groups.

To open the Equality of Two Proportions dialog box, from the menus choose:

Utilities
  Power Analysis
    Equality of Two Proportions...



Proportion 1. The hypothesized proportion in the first group. This value must lie in the interval (0,1).

Proportion 2. The hypothesized proportion in the second group. This value must lie in the interval (0,1) and cannot equal Proportion 1.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Sample sizes. You must identify how the total number of cases is distributed across the two groups:

- **Equal.** The number of cases in the first group equals the number of cases in the second group.
- **Unequal.** The number of cases differs between the two groups. If selecting this option, enter the ratio of the group 2 sample size to the group 1 sample size. A value between 0 and 1 indicates that the second group contains fewer cases than the first

group. Values above 1 correspond to the situation in which the second group is larger.

Level of test. Specify the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the proportion for each group determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size for the first group. SYSTAT determines the size of the second group using the designated ratio of the sample sizes.

Power for a range of sample sizes. Designate a range of integer sample sizes for the first group to find the corresponding power for each. For each sample, the size of the second group depends on the designated ratio of the sample sizes.

- **Smallest sample size.** The initial sample size for which power should be computed.
- **Largest sample size.** The final sample size for which power should be computed.
- **Increment.** The number of cases to add to the first group to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

Single Correlation Coefficient

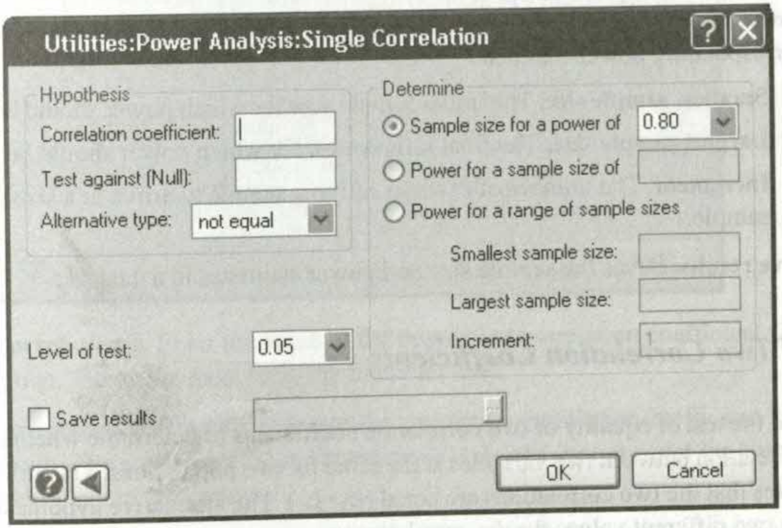
To determine whether or not two variables are related, you can test whether the population correlation coefficient (ρ) equals 0. This test assumes that you sample from a bivariate normal population and that neither of the two variables has its values fixed prior to sampling. The null hypothesis suggests that the two variables are unrelated ($\rho = 0$).

As an alternative to testing for the presence of a relationship, you can also test for a specific relationship strength. For instance, previous use of a particular psychological test yields a split-half reliability of 0.60, leading to a null hypothesis of: $\rho = 0.60$. We revise the test in an effort to improve the reliability and expect the reliability for the new test to be 0.70. Consequently, the alternative hypothesis states: $\rho = 0.70$. Notice that although we are dealing with two values of the population correlation, we are still dealing with a single sample. The correlation of 0.60 is a known standard for the

current test. In our effort to study the reliability, we will administer only the revised test to our sample.

To open the Single Correlation Coefficient dialog box, from the menus choose:

Utilities
  Power Analysis
    Single Correlation...



Correlation coefficient. Enter the value of the population correlation coefficient corresponding to the alternative hypothesis. The value must lie in the interval (-1,1).

Test against (Null). Enter the value of the population correlation coefficient corresponding to the null hypothesis. The value must lie in the interval (-1,1). Tests in which the null and alternative values are close to each other will result in very large samples, and thus may require significant computation time.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Level of test. Specify the probability of a Type I error, commonly referred to as the alpha (α) level.

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size for either group. SYSTAT assumes the two samples have the same size.

Power for a range of sample sizes. Designate a range of integer sample sizes for either group to find the corresponding power for each. The size of the first group equals the size of the second group for each sample.

- **Smallest sample size.** The initial sample size for which power should be computed.
- **Largest sample size.** The final sample size for which power should be computed.
- **Increment.** The number of cases to add to a sample to arrive at a size for the next sample.

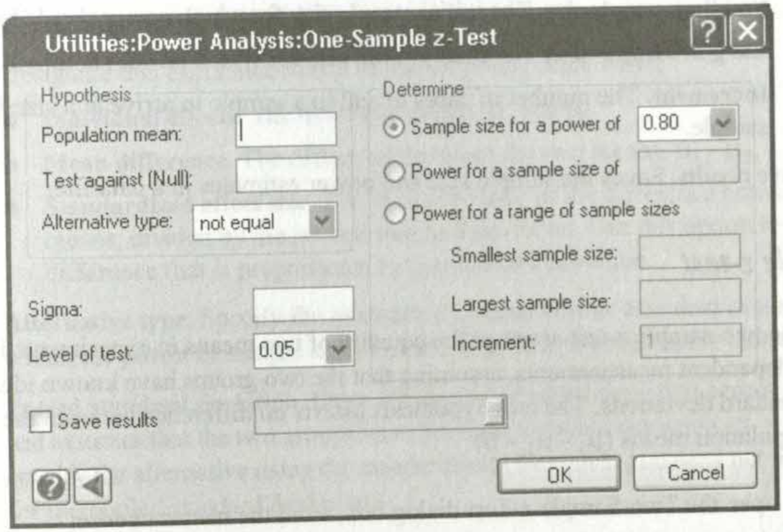
Save results. Saves the sample size and power estimates in a data file.

One-Sample z-test

The one-sample z-test compares the mean for a single population to a known standard when the population standard deviation is known. In this case, the null hypotheses specifies the known mean. The alternative hypothesis states the mean value if the population actually differs from the standard.

To open the One-Sample z-Test dialog box, from the menus choose:

Utilities
  Power Analysis
    One-Sample z-Test...



Population mean. Enter the hypothesized value of the population mean according to the alternative hypothesis.

Test against (Null). Enter the hypothesized value of the population mean according to the null hypothesis.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Sigma. State the known population standard deviation.

Level of test. Specify the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the mean, the constant against which to test the mean, and the standard deviation determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The initial sample size for which power should be computed.
- **Largest sample size.** The final sample size for which power should be computed.
- **Increment.** The number of cases to add to a sample to arrive at a size for the next sample.

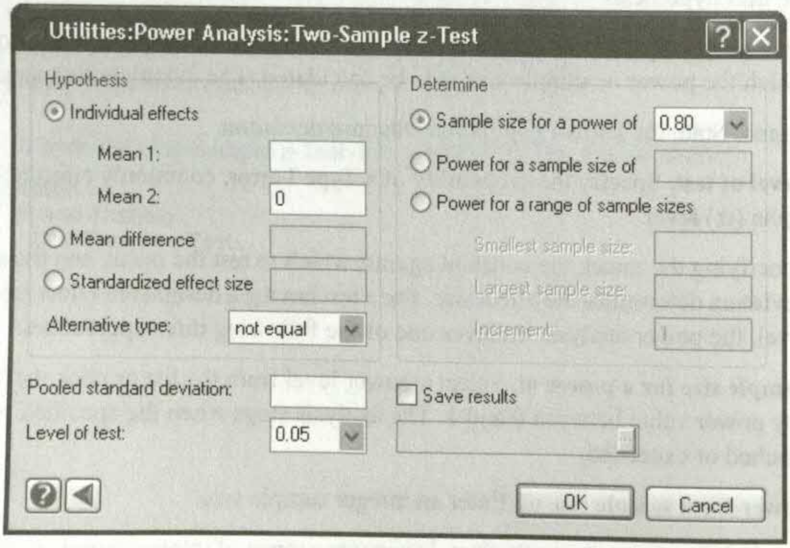
Save results. Saves the sample size and power estimates in a data file.

Two-Sample z-test

The two-sample z-test assesses the equality of two means in experiments involving independent measurements, assuming that the two groups have known identical standard deviations. The null hypothesis asserts no difference between the two population means ($\mu_1 - \mu_2 = 0$).

To open the Two-Sample z-Test dialog box, from the menus choose:

Utilities
  Power Analysis
    Two-Sample z-Test...



Hypothesis. The alternative hypothesis proposes a nonzero difference between the group means:

H_1: $\mu_1 - \mu_2 = c$, where $c < 0$ or $c > 0$ or $c \neq 0$

Designate this difference in one of the following three ways:

- **Individual effects.** The individual means, μ_1 and μ_2.
- **Mean difference.** The difference between the two means, $\mu_1 - \mu_2$.
- **Standardized effect size.** The absolute value of the difference between the two means, divided by the pooled standard deviation. Use this option to state a mean difference that is proportional to the standard deviation.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Pooled standard deviation. Enter the pooled standard deviation across groups. This test assumes that the two groups have the same standard deviation, $\sigma_1 = \sigma_2 = \sigma$. If you specify the alternative using the standardized effect size, you need not provide a value for the pooled standard deviation.

Level of test. Set the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the mean difference and the pooled standard deviation determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size. The sample size corresponds to the size of either group.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The initial sample size for either group.
- **Largest sample size.** The final sample size for either group.
- **Increment.** The number of cases to add to each group to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

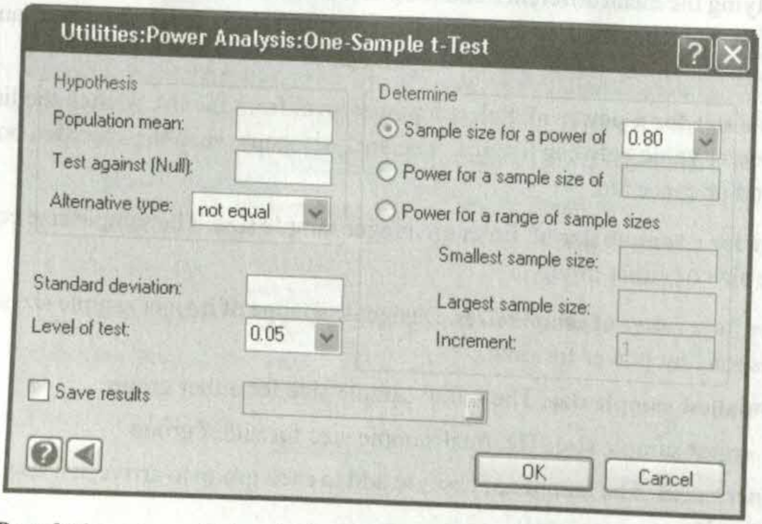
Note: Although the two-sample z-test allows different sample sizes across the two groups, the power calculations for this test assume that each group contains the same number of cases.

One-Sample t-test

Use the one-sample t-test to compare the mean for a single population to a known standard. For example, many intelligence tests have been standardized to have a mean of 100. We can administer one such test to a group in an effort to see if the group scores better or worse than the general population. In this case, the null hypotheses states a mean of 100 for the group. The alternative hypothesis states the mean value if the group actually differs from the population.

To open the One-Sample t-Test dialog box, from the menus choose:

Utilities
 Power Analysis
  One-Sample t-Test...



Population mean. Enter the hypothesized value of the population mean according to the alternative hypothesis.

Test against (Null). Enter the hypothesized value of the population mean according to the null hypothesis.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Standard deviation. State the hypothesized population standard deviation.

Level of test. Specify the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the mean, the constant against which to test the mean, and the standard deviation determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The initial sample size for which power should be computed.
- **Largest sample size.** The final sample size for which power should be computed.
- **Increment.** The number of cases to add to a sample to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

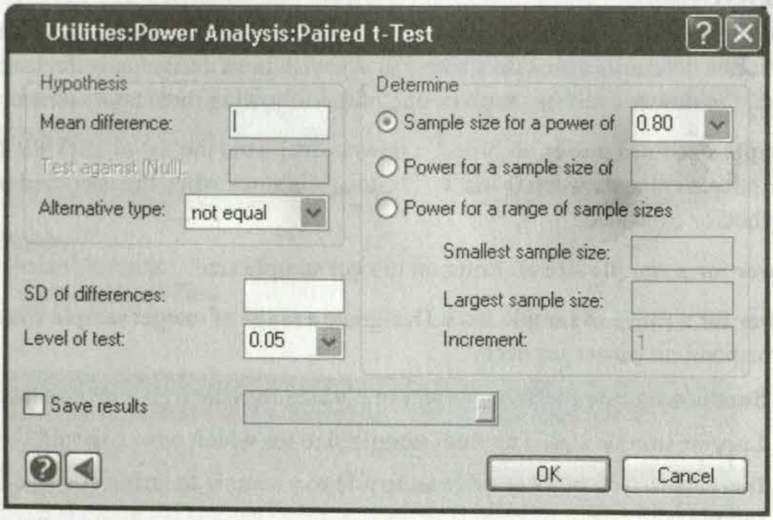
Paired t-test

The paired t-test assesses the equality of two means in experiments involving paired measurements. In practice, this test computes the differences between values of the two variables for each case and tests whether the average difference differs from 0. For instance, in a study on high blood pressure, a researcher plans to measure all patients at the beginning of the study, administer a treatment, and measure the patients again. Thus, each patient yields two measures, often called before and after measures. Subtracting the after measurement from the before measurement results in a single score for each patient that reflects the change due to the intervening treatment. The null

hypothesis asserts that the blood pressure will be the same in the two groups ($\mu_1 - \mu_2 = 0$).

To open the Paired t-Test dialog box, from the menus choose:

Utilities
  Power Analysis
    Paired t-Test...



Mean difference. Specify the expected difference between the two responses. This value corresponds to the difference proposed by the alternative hypothesis.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

SD of differences. State the standard deviation of the differences between the two responses. If you have some knowledge of the variances of the individual responses but not of the variance of their difference, estimate the correlation of the response (ρ) and use:

$$\sigma_{x-y} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y}$$

or, if the variances of the two are equal, use:

$$\sigma_{x-y} = \sqrt{2\sigma_x^2(1-\rho_{xy})}$$

Level of test. Set the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the mean difference and the standard deviation of the differences determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size. The sample size corresponds to the number of pairs for test.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The number of pairs for the initial sample.
- **Largest sample size.** The number of pairs for the final sample.
- **Increment.** The number of pairs to add to a sample to arrive at a size for the next sample.

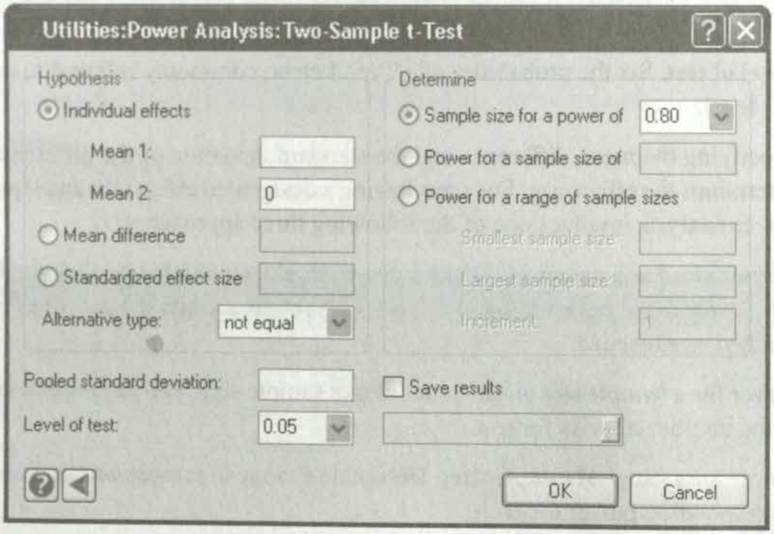
Save results. Saves the sample size and power estimates in a data file.

Two-Sample t-test

The two-sample t-test assesses the equality of two means in experiments involving independent measurements, assuming that the two groups have identical (but unknown) standard deviations. For example, suppose we wish to investigate the effects of a new drug on blood pressure. We plan to randomly assign patients with high blood pressure to one of two groups, a placebo group receiving a sugar pill and a treatment group receiving the new drug. After several months of treatments, we record the blood pressure for every member of each group. The null hypothesis suggests that the average blood pressure will not differ between the two groups ($\mu_1 - \mu_2 = 0$).

To open the Two-Sample t-Test dialog box, from the menus choose:

Utilities
  Power Analysis
    Two-Sample t-Test...



Hypothesis. The alternative hypothesis proposes a nonzero difference between the group means:

H_1: $\mu_1 - \mu_2 = c$, where $c<0$ or $c>0$ or $c \neq 0$

Designate this difference in one of the following three ways:

- **Individual effects.** The individual means, μ_1 and μ_2.
- **Mean difference.** The difference between the two means, $\mu_1 - \mu_2$.
- **Standardized effect size.** The absolute value of the difference between the two means, divided by the pooled standard deviation. Use this option to state a mean difference that is proportional to the standard deviation.

Alternative type. Specify the alternative (greater than or less than or not equal) under which the power or sample size is to be calculated. The default is 'not equal'.

Pooled standard deviation. Enter the pooled standard deviation across groups. This test assumes that the two groups have the same standard deviation, $\sigma_1 = \sigma_2 = \sigma$. If you specify the alternative using the standardized effect size, you need not provide the standard deviation.

Level of test. Set the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the mean difference and the pooled standard deviation determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size. The sample size corresponds to the size of either group.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The initial sample size for either group.
- **Largest sample size**. The final sample size for either group.
- **Increment.** The number of cases to add to each group to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

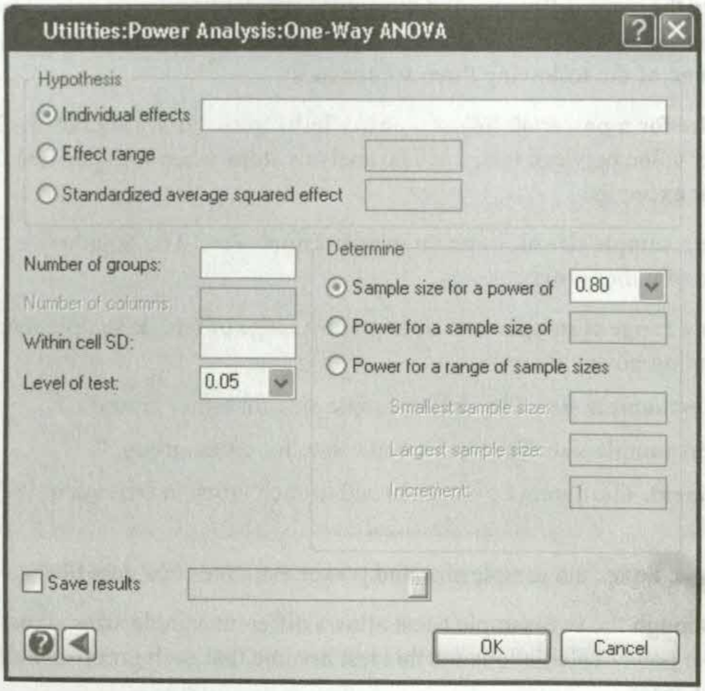
Note: Although the two-sample t-test allows different sample sizes across the two groups, the power calculations for this test assume that each group contains the same number of cases.

One-Way ANOVA

One-way analysis of variance tests the equality of two or more means in experiments involving one continuous dependent variable and one categorical independent variable, or factor. For instance, you could set up an experiment in which you record the amount of fat absorbed by doughnuts cooked in peanut oil, corn oil, and lard. The null hypothesis states that the amount of fat absorbed does not depend on the substance used in the cooking process ($\mu_1 = \mu_2 = \mu_3$).

To open the One-Way ANOVA dialog box, from the menus choose:

Utilities
  Power Analysis
    One-Way ANOVA...



Hypothesis. The alternative hypothesis asserts that at least one of the groups differs from the others. Designate this difference in one of the following three ways:

- **Individual effects.** The mean for each group, separated by commas. The number of means must equal the number of groups. SYSTAT centers the means about zero by subtracting the overall mean.
- **Effect range.** The difference between the largest mean and the smallest mean. SYSTAT generates uniformly spaced effects centered about zero covering this range.
- **Standardized average squared effect.** The average squared effect divided by the variance of a group. Use this option to state an average squared effect that is proportional to the variance within each cell.

Number of groups. Enter the number of distinct levels of the independent variable.

Within cell SD. State the hypothesized standard deviation within each group. One-way analysis of variance assumes that the standard deviation does not vary across groups.

Level of test. Set the probability of a Type I error, commonly referred to as the alpha (α) level.

Specifying the effects and the standard deviation within each group determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size. The sample size corresponds to the number of cases in each group.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The number of cases in each group for the initial sample.
- **Largest sample size.** The number of cases in each group for the final sample.
- **Increment.** The number of cases to add to each group to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

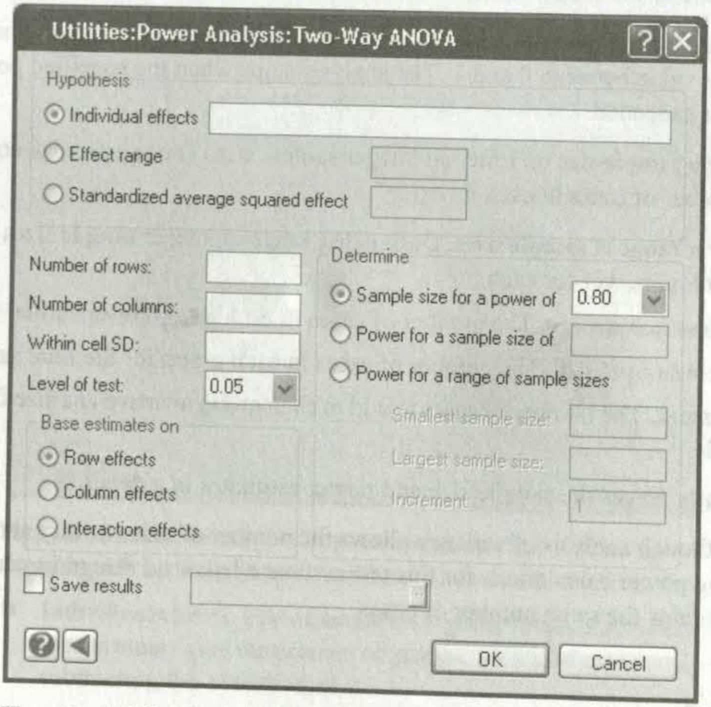
Note: Although analysis of variance allows the number of cases to vary across groups (cells), the power calculations for this test assume a balanced design in which each group contains the same number of cases.

Two-Way ANOVA

Two-way analysis of variance tests the equality of two or more means in experiments involving one continuous dependent variable and two categorical factors. As a result, the analysis involves three sets of null and alternative hypotheses. One set focuses on the row factor effect, one set focuses on the column factor effect, and the final set focuses on the interaction effect. For example, we plan to study the factors affecting the time needed to complete the Chicago marathon. For several years, we will record the time in which each runner finishes, their gender, and the weather (cold, pleasant, or hot). We can examine if males differ from females, if weather affects the completion time, and if weather affects the time differently for males and females. For experiments involving three or more factors, use generic power analysis.

To open the Two-Way ANOVA dialog box, from the menus choose:

Utilities
 Power Analysis
  Two-Way ANOVA...



Hypothesis. The alternative hypothesis states values for the row, column, or interaction effects. Designate these values in one of the following three ways:

- **Individual effects.** The effect for each group, separated by commas. SYSTAT centers the effects about zero by subtracting their mean. The number of effects to enter depends on whether the power estimates are based on the row, column, or interaction effects.
- **Effect range.** The difference between the largest effect and the smallest effect. SYSTAT generates uniformly spaced effects centered about zero covering this range.

- **Standardized average squared effect.** The average squared effect divided by the variance of a group. Use this option to state an average squared effect that is proportional to the variance within each cell.

Number of rows. Enter the number of distinct levels of the row factor.

Number of columns. Enter the number of distinct levels of the column factor.

Within cell SD. State the hypothesized standard deviation within each cell. Two-way analysis of variance assumes that the standard deviation does not vary across cells.

Level of test. Set the probability of a Type I error, commonly referred to as the alpha (α) level.

Base estimates On. The power and sample size estimation involves one of three sets of effects: the main effect for the row factor (Row effects), the main effect for the column factor (Column effects), or the interaction effect found at each combination of the row factor and the column factor levels (Interaction effects). For example, to determine the number of cases needed to correctly identify a difference between the row means 80% of the time, set the power to .80 and base the estimates on the row effects.

Regardless of the effects on which the software bases the estimates, the analysis always includes an interaction term in the ANOVA model. Consequently, there are R*C*(n-1) degrees of freedom for error, where R and C represent the number of rows and columns respectively and n equals the number of subjects in each cell. For designs without an interaction, use generic power analysis.

Notice that the effects being specified depend on the effects on which the estimates are based. For instance, basing the estimates on the row effects and entering an effect range results in R uniformly spaced effects. Basing the estimates on the interaction effects yields R*C uniformly spaced effects covering the same effect range.

Specifying the effects and the standard deviation within each group determines the effect size. For a test having a designated effect size and alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size. The sample size corresponds to the number of cases in each combination of row and column factor levels.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The number of cases at each combination of row and column factor levels for the initial sample.
- **Largest sample size.** The number of cases at each combination of row and column factor levels for the final sample.
- **Increment.** The number of cases to add to each cell to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

Note*:* Although analysis of variance allows the number of cases to vary across cells, the power calculations for this test assume a balanced design in which each cell contains the same number of cases.

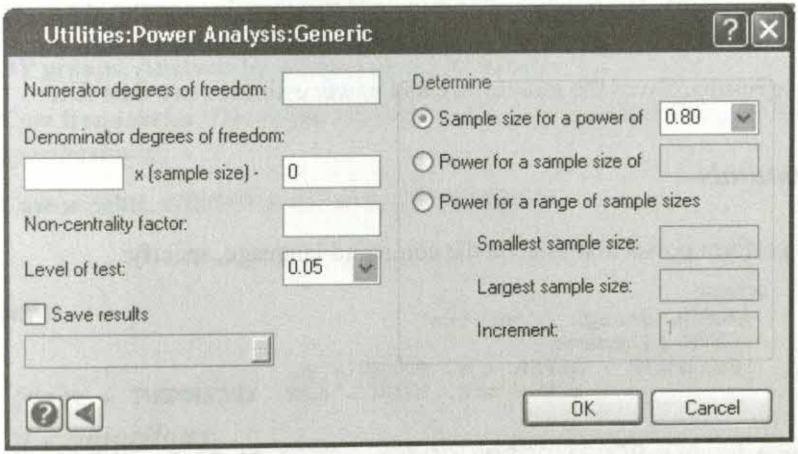
Generic Power Analysis

Generic power analysis handles any problem for which the power at the alternative is given by a non-central F-distribution meeting the following three conditions:

- fixed numerator degrees of freedom.
- denominator degrees of freedom that are linear in the sample size.
- a non-centrality parameter that is a multiple of the sample size.

To open the Generic dialog box, from the menus choose:

Utilities
  Power Analysis
    Generic...



Numerator degrees of freedom. The degrees of freedom for the effect represented in the numerator of the *F-ratio*.

Denominator degrees of freedom. The degrees of freedom for the effect represented in the denominator of the *F-ratio* must be a linear function of the sample size. State the multiplicative and additive constants defining this function.

Non-centrality factor. Specify the multiplicative constant for the sample size that yields the non-centrality parameter for the desired non-central F-distribution.

Level of test. Set the probability of a Type I error, commonly referred to as the alpha (α) level.

For a test having a designated alpha level, the power analysis involves one of the following three approaches:

Sample size for a power of. Select a power level from the list or click the list to type any power value between 0 and 1. The analysis stops when the specified power is reached or exceeded.

Power for a sample size of. Enter an integer sample size.

Power for a range of sample sizes. Designate a range of integer sample sizes to find the corresponding power for each.

- **Smallest sample size.** The initial sample size for which power should be computed.
- **Largest sample size.** The final sample size for which power should be computed.

- **Increment.** The number of cases to add to a sample to arrive at a size for the next sample.

Save results. Saves the sample size and power estimates in a data file.

Using Commands

To perform power analysis via the command language, specify:

```
POWER
  MODEL design / options
  SAVE filename
  ESTIMATE / ALPHA = a, POWER = p,
             LOW = min, HIGH = max, INCREMENT = step
```

where design is PROP1, PROP2, CORR1, CORR2, Z1, Z2, T1, PAIRED, T2, ONEWAY, TWOWAY, or GENERIC. The model options depend on the design.

```
MODEL PROP1 / P1 = p1,  NULL = p0, ALTER= lt/gt/ne
MODEL PROP2 / P1 = p1, P2 = p2, RATIO =r,ALTER= lt/gt/ne
MODEL CORR1 / COEF1 = c1,  NULL = c0,ALTER= lt/gt/ne
MODEL CORR2 / COEF1 = c1,  COEF2 = c2,  N1 = n1,  N2 = n2
ALTER= lt/gt/ne
MODEL Z1 / M1 = m1,  NULL = m0,  SIGMA = s, ALTER= lt/gt/ne
MODEL Z2 / M1 = m1,  M2 = m2,  RANGE = d,
                STEFF = g,  POOLED = s,ALTER= lt/gt/ne
MODEL T1      / M1 = m1,  NULL = m0,  SD = s,ALTER= lt/gt/ne
MODEL PAIRED  / DIFF = d,  WITHIN = s,ALTER= lt/gt/ne
MODEL T2      / M1 = m1,  M2 = m2,  RANGE = d,
                STEFF = g,  WITHIN = s,ALTER= lt/gt/ne
MODEL ONEWAY      / GROUPS = k,  EFFECT = m1,m2,...,mk,
                    RANGE = d,  AVGESQ = g,  WITHIN = s
MODEL TWOWAY      / ROWS = r,  COLUMNS = c,
                    EFFECT = m1,m2,...,mrxc,
                    RANGE = d,  AVGESQ = g,  WITHIN = s
                    REFFECTS or CEFFECTS or IEFFECTS
MODEL GENERIC     / NDF = df,  C1 = m,  C0 = b,  NCP = a
```

Usage Considerations

Types of data. No data file is needed for power analysis.

Print options. The output is standard for all PLENGTH options. Use PLENGTH NONE to suppress the output display.

Quick Graphs. The Quick Graph displays the power curve (power versus sample size) corresponding to the designated alpha and effect size.

Saving files. POWER saves the sample size estimates with the corresponding power.

BY groups. Analysis by groups is not available.

Case frequencies. The power calculations ignore any FREQUENCY variable specifications.

Case weights. WEIGHT is not available in POWER.

Examples

Example 1
Equality of Proportions

Cohen (1988) presents a situation in which a psychologist wishes to study the effects of birth order in anxiety producing conditions. Previous research in the United States suggests that when expecting an anxiety-inducing situation, two-thirds of first-born and people without siblings prefer companionship while waiting. In contrast, one-half of those who have siblings but are not the eldest prefer waiting alone.

	Companionship	Alone
Eldest or only child	0.667	0.333
Not the eldest child	0.500	0.500

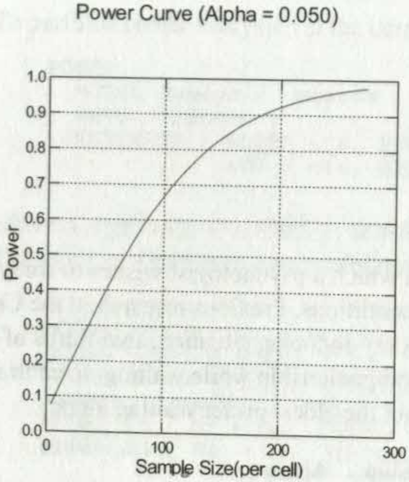
To test for a similar difference in preference in another country, the researcher obtains 80 people in each birth order condition. Using an alpha level of 0.05, you can determine the power using the following commands:

```
POWER
   MODEL PROP2 / P1=0.667 P2=0.5
   ESTIMATE / ALPHA=0.05 LOW=80
```

The output is:

```
Test for Equality of Proportions with Alternative 'not equal'

   P1              : 0.667
   P2              : 0.500
   Ratio(2:1)      : 1.000
   Effect Size     : 0.341 Approximate Test
   Effect Size     : 0.167 Large Sample Test
   ALPHA           : 0.050
```

```
Sample Size:  Low  :      80
Sample Size:  High :      80
Increment          :       1

      Approximate Test             Large Sample Test
Group1   Group2    POWER   Group1   Group2    POWER
-----------------------------------------------------
    80       80    0.577      80       80     0.572

Total Sample Size = 160            Total Sample Size = 160
```



Note that the power reported by both approximate and large sample test is almost the same.

One-sided vs. Two-sided Test

To see the difference in the power curve of a one-sided and two-sided test, we begin by generating and saving the power estimates at sample sizes ranging from 20 to 200 for an alpha of 0.05.

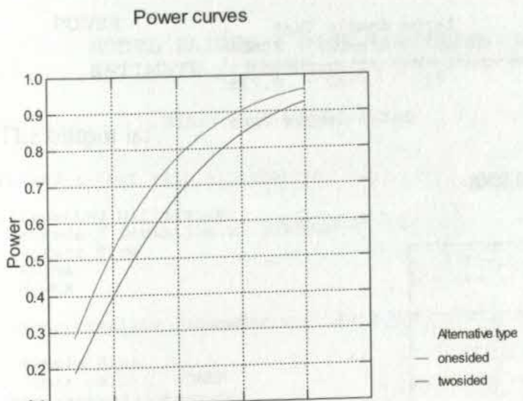
```
POWER
  MODEL PROP2/ P1=0.667 P2=0.5
  SAVE TWOSIDED
  EST/LOW=20 HIGH=200
  USE TWOSIDED
  LET ALTER$='twosided'
  ESAVE TWOSIDED
```

We now repeat the same model with one- sided alternative by using ALTER command.

```
POWER
  MODEL PROP2/ P1=0.667 P2=0.5 ALT=GT
  SAVE ONESIDED
  EST/LOW=20 HIGH=200
  USE ONESIDED
  LET ALTER$='onesided'
  ESAVE ONESIDED
```

The saved file reports power by both approximate (*AT_POWER*) and large sample (*LST_POWER*) tests. We plot the power curves using power by approximate test against sample size.

```
APPEND TWOSIDED ONESIDED
LET SIZE=AT_GROUP1
LET POWER=AT_POWER
PLOT POWER*SIZE/GROUP=ALTER$ OVERLAY SMOOTH=SPLINE XGRID YGRID,
               SIZE=0.0, 0.0,XLAB=' Sample Size (per group)',
               YLABEL='Power', LTITLE='Alternative type',
               TITLE='Power curves'
```



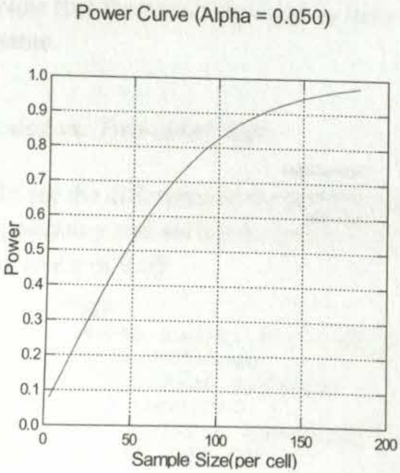
Unequal Group Sizes

Suppose that instead of 80 people in each birth order category, we have access to three times as many people who are not the oldest child. The RATIO option defines the ratio of the sample size for the second group to the sample size of the first group. Using an alpha of 0.05, we can find the power for the test using:

```
POWER
   MODEL PROP2 / P1=0.667 P2=0.5 RATIO=3
   ESTIMATE / ALPHA=0.05 LOW=80
```

The output is:

```
Test for Equality of Proportions with Alternative 'not equal'

  P1                      : 0.667
  P2                      : 0.500
  Ratio(2:1)              : 3.000
  Effect Size             : 0.341 Approximate Test
  Effect Size             : 0.167 Large Sample Test
  ALPHA                   : 0.050
  Sample Size:  Low       :    80
  Sample Size:  High      :    80
  Increment               :     1

      Approximate Test            Large Sample Test
  Group1   Group2    POWER    Group1    Group2    POWER
 ------------------------------------------------------
    80       240     0.751      80        240     0.738

Total Sample Size = 320              Total Sample Size = 320
```



The power of the test increases from 0.577 to 0.751 when the second group contains three times as many people as the first group.

Example 2 *Paired t-Test*

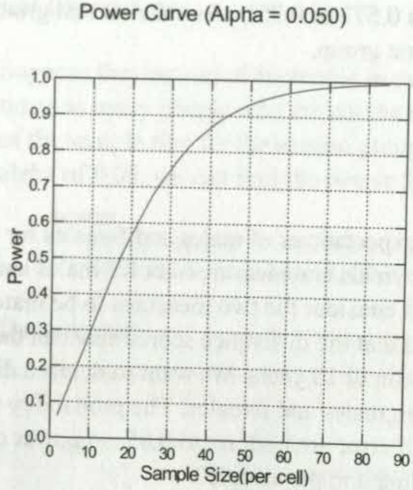
Suppose we have access to the life expectancies of males and females for countries throughout the world. Each country yields one measurement for males and one measurement for females, so we can consider the two measures to be matched by country. Past research has suggested that the difference scores between the two life expectancies have a standard deviation of 10 years. We want to detect a difference of five years in life expectancy between males and females. The probability of a Type I error and the probability of a Type II error are both set to 0.05, so power equals 0.95. How many countries should be included in the sample?

The input is:

```
POWER
   MODEL PAIRED / DIFF=5 WITHIN=10
   ESTIMATE / ALPHA=0.05 POWER=0.95
```

The output is:

```
Paired t-test with Alternative 'not equal'

  Expected Difference                  :  5.000
  Standard Deviation of Difference     : 10.000
  Effect Size                          :  0.500
  ALPHA                                :  0.050
  POWER                                :  0.950

  Noncentrality Parameter = 0.500 * sqrt(sample size)

  Sample Size
   (per cell)   POWER
  ------------------
          50    0.934
          51    0.938
          52    0.943
          53    0.947
          54    0.950

  Total Sample Size = 54 Pairs
```



We should sample 54 countries to achieve a power of 0.95.

Overlaying Power Curves

The Quick Graph for power analysis displays the relationship between power and the sample size for a fixed alpha level and effect size. A different alpha level results in a different power curve. In this example, we demonstrate how to generate a single plot displaying the power curves for multiple alpha levels for a paired t-test. This plot allows you to examine the relationship between power, sample size, and alpha for a fixed effect size.

We begin by generating and saving power estimates at sample sizes ranging from 5 to 80 for an alpha of 0.05. Because we plan to overlay multiple curves in a single plot, we open the saved estimates and create a new variable, *ALPHA*, to denote the alpha level used.

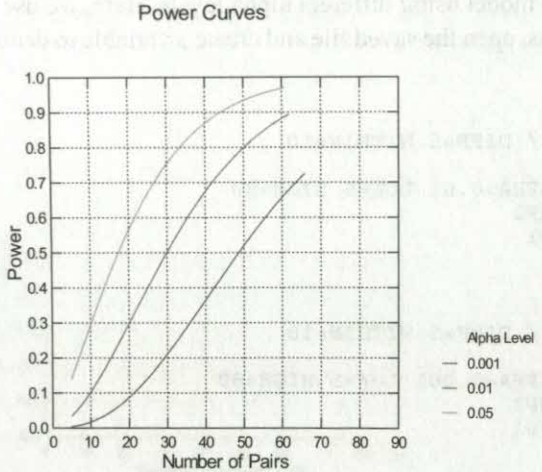
```
POWER
  GRAPH NONE
  MODEL PAIRED / DIFF=5 WITHIN=10
  SAVE CURVES
  ESTIMATE / ALPHA=0.05 LOW=5 HIGH=80
  USE CURVES
  LET ALPHA=0.05
  ESAVE CURVES
```

We then repeat the same model using different alpha levels. Here, we use 0.01 and 0.001. After each analysis, open the saved file and create a variable to denote the alpha level.

```
POWER
  MODEL PAIRED / DIFF=5 WITHIN=10
  WORK TEMP1
  ESTIMATE / ALPHA=0.01 LOW=5 HIGH=80
  USE &WORK\TEMP1
  LET ALPHA=0.01
  EWORK TEMP1

POWER
  MODEL PAIRED / DIFF=5 WITHIN=10
  WORK TEMP2
  ESTIMATE / ALPHA=0.001 LOW=5 HIGH=80
  USE &WORK\TEMP2
  LET ALPHA=0.001
  EWORK TEMP2
  GRAPH
```

We now have three files (*CURVES.SYZ, TEMP1.SYZ*, and *TEMP2.SYZ*) containing power values for samples ranging in size from 5 to 80. We append these files together to generate a single file containing all of the information needed for the plot and create the plot using the following commands:

```
APPEND CURVES &WORK\TEMP1
ESAVE CURVES
APPEND CURVES &WORK\TEMP2
PLOT POWER*SIZE / OVERLAY GROUP=ALPHA SMOOTH=SPLINE,
                  SHORT XLABEL='Number of Pairs' XGRID,
                  YLABEL='Power' YGRID SIZE=0,0,0 COLOR=2,1,3,
                  LTITLE='Alpha Level' TITLE='Power Curves'
```



In this single plot, we can see that regardless of the alpha level, as sample size increases, power increases. However, for any given power, the smaller the alpha level, the more the cases that are needed to yield that power.

This technique of appending saved estimates from multiple power analyses can also be generalized to study how effect size relates to power. Vary the standard deviation or mean difference, create a variable in the output data set to denote the effect size, append the output files, and use the effect size variable as the grouping variable.

Power Surfaces

Overlaying power curves is one technique for viewing the relationship between three of the underlying components of power analysis while holding the fourth constant. However, the overlay plots can become difficult to use when several curves appear. A power surface alleviates this problem by spacing the individual power curves along a third dimension and fitting a surface to the curves.

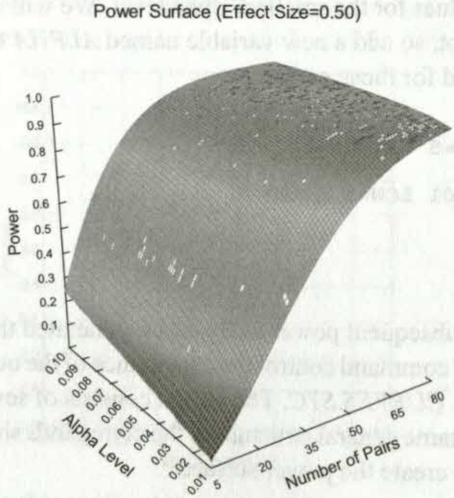
In this example, we fix the effect size of the paired t-test at a constant and study the relationship between alpha, sample size, and power at this effect size. Because we will be generating several power curves, we use FOR...NEXT looping and printing to generate the commands for all of the analyses and submit the resulting file to create the data file needed for the plot.

Begin by saving the power values for the smallest alpha level. We will use alpha as the Y-variable for the surface plot, so add a new variable named *ALPHA* to the saved file to denote the alpha level used for those estimates.

```
POWER
  MODEL PAIRED / DIFF=5 WITHIN=10
  SAVE ALPHA
  ESTIMATE / ALPHA=0.01 LOW=5 HIGH=80
  USE ALPHA
  LET ALPHA=0.01
  ESAVE ALPHA
```

The commands needed for the subsequent power analysis are generated through FOR...NEXT loops. The FORMAT command control the appearance of the output, which gets sent to an ASCII file named *CURVES.SYC*. The output consists of several sets of SYSTAT commands having the same general structure as the commands shown above. We plot the resulting data file to create the power surface.

```
CLASSIC ON
FORMAT 5,2
PUSH CLASSIC
OUTPUT CURVES.SYC / NOSCREEN
FOR i=0.02 TO 0.10 STEP 0.01
PRINT "PUSH GRAPH"
PRINT "GRAPH NONE"
PRINT "OUTPUT / NOSCREEN"
PRINT "POWER"
PRINT "MODEL PAIRED / DIFF=5 WITHIN=10"
PRINT "WORK 'ALOOP'"
PRINT "ESTIMATE / ALPHA=",i, "LOW=5 HIGH=80"
PRINT "USE '&WORK\ALOOP'"
PRINT "LET ALPHA=",i
PRINT "EWORK ALOOP"
PRINT "NEW"
PRINT "APPEND ALPHA '&WORK\ALOOP'"
PRINT "ESAVE ALPHA"
PRINT " "
PRINT "POP GRAPH"
PRINT "OUTPUT"
NEXT
OUTPUT
SUBMIT '&OUTPUT\CURVES'
POP CLASSIC

PLOT POWER*ALPHA*SIZE / XMIN=5 XMAX=80 YMIN=0.01 YMAX=0.10,
                        SMOOTH=NEXPO SURFACE=COLOR CUT=80,
                        XLAB='Number of Pairs' YLAB='Alpha Level',
                        ZLAB='Power' SIZE=0 XTICK=5 YTICK=9,
                        TITLE='Power Surface (Effect Size=0.50)'
```



In this graph, we can see the relationship between power, alpha, and sample size for any alpha level between 0.01 and 0.10 and for any sample size between 5 and 80. We are no longer restricted to the curves for select alpha values appearing in the overlay plot.

Example 3
Independent Samples t-Test

Jaccard (2001) describes an experiment by Benson involving transcendental meditation. This relaxation technique involves repeating a nonsense, two-syllable "mantra" quietly to oneself for a period of time. Transcendental meditation instructors assign mantras to individuals, stressing the importance of the mantra itself. Benson suggests that optimal relaxation can be achieved by repetition of any sound, such as the word "one".

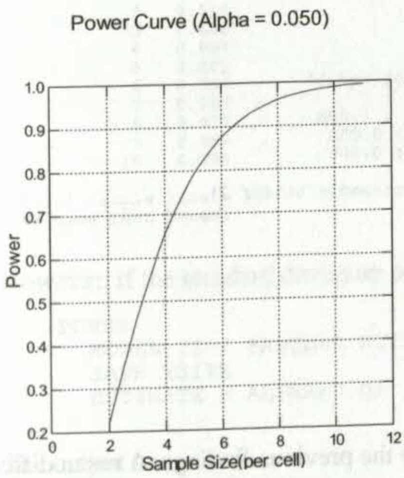
Suppose we were to conduct an experiment to compare the heart rates of individuals using standard transcendental meditation mantras and those using a mantra of "one". Previous research suggests using an estimate of the standard deviation for either group of 2. Let the probability of incorrectly rejecting the null hypothesis be 0.05. To

determine how many subjects we should employ to detect a difference of four heartbeats per minute with a power of 0.80, submit the following commands:

```
POWER
   MODEL T2 / RANGE=4 WITHIN=2
   ESTIMATE / ALPHA=0.05
```

The output is:

```
Two Sample t-test with Alternative 'not equal'

  Effect Size                 : 2.000
  Pooled Standard Deviation : 2.000
  Range                       : 4.000
  ALPHA                       : 0.050
  POWER                       : 0.800

Noncentrality Parameter = 2.000 * sqrt(sample size / 2)

  Sample Size
   (per cell)   POWER
  -------------------
          2    0.218
          3    0.463
          4    0.657
          5    0.791
          6    0.876

  Total Sample Size = 12
```



To achieve a power of at least 0.80, we should use six people in each group, for a total sample size of 12.

Standardized Effect Size

As an alternative to estimating the population standard deviation, you can define the difference in means as a multiple of the standard deviation. Suppose that we are interested in detecting a mean difference of two standard deviations.

H_0: μ_1-$\mu_2 = 0$

H_1: μ_1-$\mu_2 = 2\sigma$

This alternative hypothesis corresponds to:

$$(\mu_1\text{-}\mu_2)/\sigma = 2$$

Consequently, the standardized effect size (STEFF) equals 2. In this case, we do not need to specify the standard deviation (WITHIN). To detect this difference with a power of 0.80 ($\alpha = 0.05$), the input is:

```
POWER
   MODEL T2 / STEFF=2
   ESTIMATE / ALPHA=0.05
```

The output is:

```
Two Sample t-test with Alternative 'not equal'

  Standardized Average Squared Effect : 1.000
  ALPHA                               : 0.050
  POWER                               : 0.800

Noncentrality Parameter = 2.000 * sqrt(sample size / 2)

  Sample Size
   (per cell)   POWER
 --------------------
            2   0.218
            3   0.463
            4   0.657
            5   0.791
            6   0.876

  Total Sample Size = 12
```

Notice that the results are identical to the previous findings. A mean difference of 4 with a standard deviation of 2 equals a standardized mean difference of 2.

t-Tests vs. z-Tests

If we can safely assume that the population standard deviation is actually known, we can use the z-test to test our hypothesis of no difference between the two groups. To determine the power for samples ranging in size from 2 to 10 per group, the command input is:

```
POWER
   MODEL Z2 / RANGE=4 POOLED=2
   SAVE ZSIZE
   ESTIMATE / ALPHA=0.05 LOW=2 HIGH=10
```

The output is:

```
Two Sample z-test with Alternative 'not equal'

 Mean Difference             : 4.000
 Pooled Standard Deviation   : 2.000
 Effect Size                 : 2.000
 ALPHA                       : 0.050
 Sample Size:  Low           :     2
 Sample Size:  High          :    10
 Increment                   :     1

 Sample Size
  (per cell)   POWER
--------------------
          2   0.516
          3   0.688
          4   0.807
          5   0.885
          6   0.934
          7   0.963
          8   0.979
          9   0.989
         10   0.994

 Total Sample Size = 20
Data have been Saved.
```

However, if the standard deviation must be estimated, use the t-test.

```
POWER
   MODEL T2 / RANGE=4 WITHIN=2
   SAVE TSIZE
   ESTIMATE / ALPHA=0.05 LOW=2 HIGH=10
```

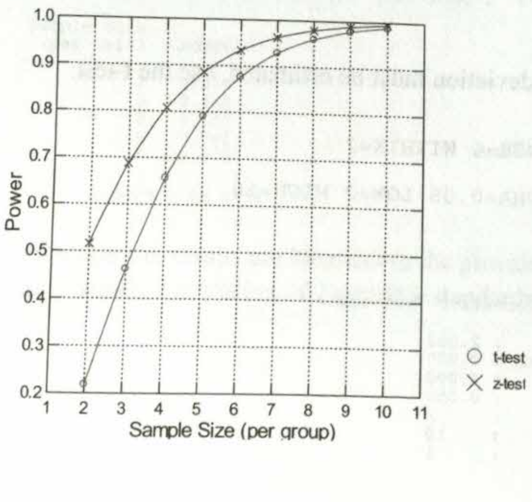
The output is:

```
Two Sample t-test with Alternative 'not equal'

 Effect Size                 : 2.000
 Pooled Standard Deviation   : 2.000
 Range                       : 4.000
 ALPHA                       : 0.050
 Sample Size:  Low           :     2
 Sample Size:  High          :    10
 Increment                   :     1
```

```
Noncentrality Parameter = 2.000 * sqrt(sample size / 2)

  Sample Size
   (per cell)    POWER
 -------------------------
           2    0.218
           3    0.463
           4    0.657
           5    0.791
           6    0.876
           7    0.929
           8    0.960
           9    0.978
          10    0.988

Data have been Saved.
```

Combining the power estimates from the two tests in a single file allows us to compare the results. Before appending the files together, we create a variable named *TEST$* to denote the test that produced the estimates. Because the t-test saves the sample size estimates as *GROUP1* and *GROUP2* instead of using the variable *SIZE* for the sample size of either group, we also create *SIZE* to simplify combining the files.

```
USE ZSIZE
LET TEST$='z-test'
ESAVE ZSIZE
USE TSIZE
LET SIZE=GROUP1
LET TEST$='t-test'
ESAVE TSIZE

APPEND TSIZE ZSIZE / INTERSECTION
PLOT POWER*SIZE / GROUP=TEST$ OVERLAY SMOOTH=SPLINE,
                  XLAB='Sample Size (per group)' YLAB='Power',
                  XGRID YGRID LTITLE=NONE
```



For any power level, the number of subjects needed for the z-test is smaller than the number needed for the t-test. Moreover, for any given sample size, the z-test yields a higher power than the t-test. As the sample gets bigger, however, the difference between the two curves decreases due to the fact that as sample size increases, the t-distribution approaches the normal distribution.

Example 4
One-Way ANOVA and Sample Size Estimation

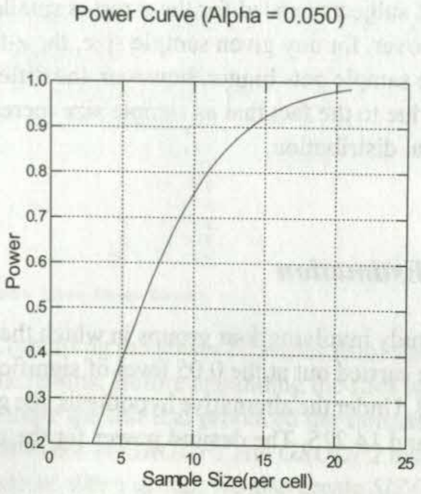
Fleiss (1986) discusses a parallel study involving four groups in which the overall test of equality of group means is to be carried out at the 0.05 level of significance. The within group standard deviation is 3. Under the alternative hypothesis, the group means are estimated to be 9.775, 12, 12, and 14.225. The desired power for the test is 0.80.

The input is:

```
POWER
  MODEL ONEWAY / GROUPS=4 EFFECT=9.775,12,12,14.225,WITHIN=3
  ESTIMATE / ALPHA=0.05 POWER=0.8
```

The output is:

```
One-way ANOVA

  Number of Groups                    :      4
  Within Cell Standard Deviation :  3.000
  Mean(1)                             :  9.775
  Mean(2)                             : 12.000
  Mean(3)                             : 12.000
  Mean(4)                             : 14.225
  Effect Size                         :  0.524
  ALPHA                               :  0.050
  POWER                               :  0.800
Noncentrality Parameter = 1.100 * Sample Size

  Sample Size
   (per cell)    POWER
  --------------------
            2    0.113
            3    0.200
            4    0.293
            5    0.386
            6    0.476
            7    0.558
            8    0.633
            9    0.698
           10    0.755
           11    0.803

  Total Sample Size = 44
```



To achieve a power of 0.80, each group should contain 11 cases.

Generic Power Analysis for One-way ANOVA

In one-way analysis of variance, if the population means actually differ, the *F-ratio* follows a non-central F-distribution. The linear relationship between the sample size and the degrees of freedom for the denominator allow us to use generic power analysis to determine sample size requirements or power values. This approach requires an input reflecting the degrees of freedom (NDF, C1, C0) and the non-centrality parameter (NCP) as follows:

- In one-way analysis of variance, the numerator degrees of freedom equal the number of groups minus 1. Here, the number of groups is 4, so NDF = 3.
- Generic power analysis requires that the denominator degrees of freedom and the number of subjects in each cell have a linear relationship. For one-way analysis of variance, the denominator degrees of freedom equal the total number of cases minus the number of groups. In this example, the total number of cases is 4*n, so the degrees of freedom equal 4n – 4. Consequently, C1 = 4 and C0 = 4.
- The non-centrality parameter must be a function of the sample size within each cell. For one-way analysis of variance, the non-centrality parameter equals the number of cases within each cell times the sum of the squared effects divided by the within cell variance.

The SYSTAT calculator assists in these computations. The command:

```
CALC AVG(9.775,12,12,14.225)
```

returns a mean of 12 for these four values. We need to subtract this value from the individual means, converting the means into effects centered at zero. We then square the effects, sum the results, and divide by the variance to derive the factor that, when multiplied by the sample size, yields the non-centrality parameter.

```
CALC (((9.775-12)^2)+((14.225-12)^2))/3^2
```

The non-centrality parameter equals the result, 1.100, times the sample size, so the non-centrality factor (NCP) is 1.100. We now have the information needed to derive the sample size needed to achieve a power of 0.80 at an alpha level of .05.

The input is:

```
POWER
   MODEL GENERIC / NDF=3 C1=4 C0=4 NCP=1.1
   ESTIMATE / ALPHA=0.05 POWER=0.8
```

The output is:

```
Generic Model

 Numerator df                 :    3
 Denominator df               :    4 * Sample Size - 4
 Noncentrality Parameter : 1.100   * Sample Size
 ALPHA                        : 0.050
 POWER                        : 0.800

 Sample Size
   (per cell)   POWER
 ----------------------
            2   0.113
            3   0.200
            4   0.293
            5   0.386
            6   0.475
            7   0.558
            8   0.633
            9   0.698
           10   0.755
           11   0.803
```

The results are identical to those found earlier.

Example 5
Two-way ANOVA

Do lesions in parts of the brain affect memory? Odeh and Fox (1991) describe a two-factor experiment by Glick and Greenstein (1973) that addresses this issue using aversive conditioning. Two groups of mice were trained to avoid an area of a two-partition box. For the first group, the researchers placed each mouse in the box and administered an electric shock when the mouse attempted to cross to the other partition. For the second group, an attempt to change partitions resulted in removal of the mouse from the box instead of a shock.

After training both groups, the researchers implanted electrodes in the hippo campus or the candate regions of the brain of each mouse. Some mice received an electrical current via the electrodes, resulting in lesions in that brain area. The remaining mice underwent the surgery with no subsequent current to cause lesions. Afterwards, Glick and Greenstein returned each mouse to the box and recorded the time to switch partitions.

This experiment corresponds to a two-way analysis of variance using lesion area and aversion technique as the two factors:

	Hippocampus Lesion	**Candate Lesion**	**No Lesion**	**Total**
Shock	*n*	*n*	*n*	*3n*
Removal	*n*	*n*	*n*	*3n*
Total	*2n*	*2n*	*2n*	*N*

where *n* represents the number of mice at each combination of lesion area and aversion technique.

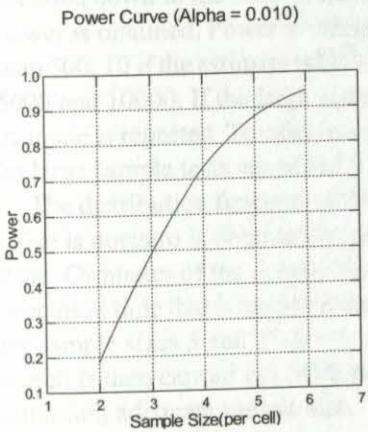
How many mice are needed to achieve a power of 0.80 at an alpha level of .01 when the average squared effect for lesion area equals two-thirds of the within-cell variance? To answer this question, submit the following commands:

```
POWER
  MODEL TWOWAY / ROWS=2 COLUMNS=3 AVGESQ=0.6667 CEFFECTS
  ESTIMATE / ALPHA=0.01
```

The CEFFECTS option identifies the standardized average squared effect as corresponding to the effects for column factor.

The output is:

```
Two-way ANOVA

 Number of Rows                          :     2
 Number of Columns                       :     3
 Standardized Average Squared Effect : 0.667
 ALPHA                                   : 0.010
 POWER                                   : 0.800

Estimate to be based on Column Main Effects.
Noncentrality Parameter = 4.000 * Sample Size

 Sample Size
  (per cell)    POWER
 -------------------
           2    0.185
           3    0.492
           4    0.733
           5    0.876
 Total Sample Size = 30
```



Four mice at each factor combination yields a power of only 0.73. To achieve a power of at least 0.80, the experiment needs five mice at each combination, corresponding to a total sample size of 30.

Interaction Effects

In determining the sample size needed to attain a specified power, we can focus on the row effects, the column effects, or the interaction effects. Previously we used CEFFECTS to define the standardized average squared effect for the column factor. In this example, we test for an effect due to the interaction between the two factors of the Glick and Greenstein brain lesion study.

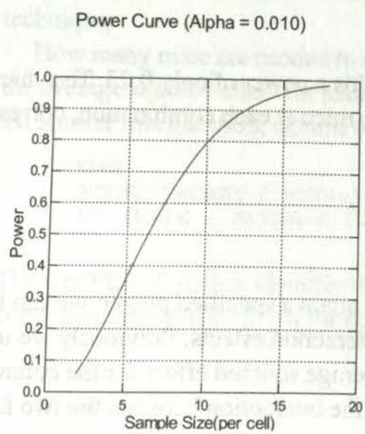
We set the alpha level to 0.01 and desire a power of 0.80. What sample size is needed to detect an average squared interaction effect that is one-fourth of the within cell variance?

The input is:

```
POWER
   MODEL TWOWAY / ROWS=2 COLUMNS=3 AVGESQ=0.25 IEFFECTS
   ESTIMATE / ALPHA=0.01 POWER=0.8
```

The output is:

```
Two-way ANOVA

  Number of Rows                             :   2
  Number of Columns                          :   3
  Standardized Average Squared Effect        : 0.250
  ALPHA                                      : 0.010
  POWER                                      : 0.800

Estimate to be based on Interaction Effects.
Noncentrality Parameter = 1.500 * Sample Size

  Sample Size
   (per cell)    POWER
  ---------------------
            2    0.059
            3    0.145
            4    0.248
            5    0.358
            6    0.467
            7    0.567
            8    0.657
            9    0.733
           10    0.795
           11    0.846

 Total Sample Size = 66
```



In this case, a total of 66 mice would be required.

Computation

Algorithms

POWER uses algorithms described by Cohen (1988). SYSTAT uses a large sample approximation based on the noncentral chi-square distribution to get a rough estimate for the necessary sample size. If the rough estimate is no greater than 40, power calculations using the noncentral F-distribution begin with the rough estimate less 10. If the rough estimate exceeds 40, the calculations begin with the rough estimate rounded down to the nearest increment. Computations continue until the required power is obtained. Power is calculated in increments of 1 if the rough estimate is less than 500, 10 if the estimate is between 500 and 5000, and 100 if the estimate is between 5000 and 10000. If the large sample estimate exceeds 10000, only the large sample estimate is reported. To calculate power and sample sizes for proportion, results based on large sample tests are added to the existing sine inverse approximate test.

The distribution function of the sample correlation coefficient when the population value is nonzero is obtained by performing numerical integration using Simpson's Rule. Ordinates of the density function are calculated recursively, resulting in an execution time that is proportional to sample size. POWER reports the power of the test for sample sizes 3 and 2^k ($k = 1, 2, ...$) until the required power is exceeded. A binary search is then carried out (with intermediate results not reported) to locate the minimum adequate sample size.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. Mahwah, N.J.: Lawrence Erlbaum.

Fleiss, J.L. (1986). *The design and analysis of clinical experiments*. New York: John Wiley & Sons.

Glick, S.D. and Greenstein, S. (1973). Comparative learning and memory deficits following hippocampal and candate lesions in mice. *Journal of Comparative and Physiological Psychology*, *82*, 188-194.

Jaccard, J. and Becker, M.A. (2001). *Statistics for the behavioral sciences*. Belmont, CA: Wadsworth Publishing Company.

Graybill, F.A. (1961). *An introduction to linear statistical models (Vol. 1)*. New York: McGraw-Hill Book Company.

Kendall, M.G. , Ord, K.J., Stuart, A. and Arnold, S. (1991). *Kendall's advanced theory of statistics*, Volume 2A. London: Hodder Arnold.

* Kraemer, H.C. and Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: SAGE Publications, Inc.

Odeh, R.E. and Fox, M. (1991). *Sample size choice: charts for experiments with linear models*. 2nd ed. Boca Raton: CRC Press.

Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley & Sons.

Tang, P.C. (1938). The power function of the analysis of variance test with tables and illustrations of their use. *Statistical Research Memoirs*, *2*, 126-149.

(*indicates additional reference.)

Chapter
3

Probability Calculator

Mangalmurti Badgujar and S. Anoopama

The Probability Calculator computes values from a probability density function, cumulative distribution function, inverse cumulative distribution function, and upper-tail probabilities for a wide variety of univariate discrete and continuous probability distributions. For continuous distributions, SYSTAT plots the graph of the probability density function and the cumulative distribution function.

Statistical Background

The statistical analyst frequently encounters probability distributions at work: in data modeling, while working with sampling distributions, in Monte Carlo simulations, and in numerous other situations. Typically, the analyst must compute values from the probability mass or density function (DF), the cumulative distribution function (CF), the inverse cumulative distribution function (IF), and upper-tail probability (1-CF). The Probability Calculator helps, for instance, to compute upper-tail and lower-tail probabilities, critical points, and *p-values* that are useful in hypothesis testing, and in constructing confidence intervals and decision rules. This tool also provides graphs of the density function and the cumulative distribution function, with areas shaded or points indicated, as a visual aid for continuous distributions.

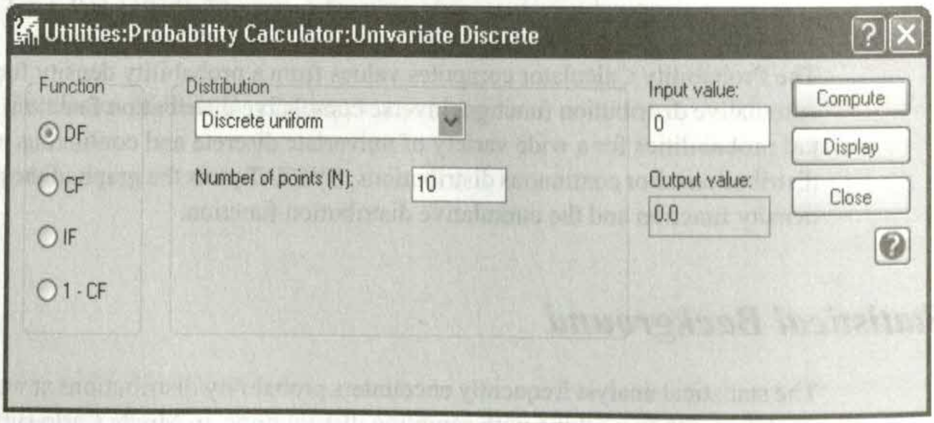
A list of distributions and expressions for these functions appears in the volume: Data: Chapter 4: Data Transformations: Distribution Functions. For more information about probability distributions and related results, see Evans et al. (2000) and Johnson et al. (1994, 1995, 2005).

Probability Calculator in SYSTAT

Univariate Discrete Distributions Dialog Box

To open the Probability Calculator: Univariate Discrete Distributions dialog box, from the menus choose:

Utilities
  Probability Calculator
    Univariate Discrete…



Function. You need to choose one of the following options:

- **DF.** Probability mass function
- **CF.** Cumulative distribution function
- **IF.** Inverse cumulative distribution function
- **1-CF.** Upper-tail probability

Distribution. From the drop-down list, select a distribution, and specify its parameters. The following choices are available:

- **Benford's law.** Benford's law with parameter *B* (base).
- **Binomial.** Binomial distribution with parameters *n* and *p*.
- **Discrete uniform.** Discrete uniform distribution with parameter *N*.
- **Geometric.** Geometric distribution with parameter *p*.
- **Hypergeometric.** Hypergeometric distribution with parameters *N*, *m*, and *n*.

- **Logarithmic series.** Logarithmic series distribution with parameter theta.
- **Negative binomial**. Negative binomial distribution with parameters k and p.
- **Poisson.** Poisson distribution with parameter lambda.
- **Zipf.** Zipf distribution with parameter a.

Input value. Enter an appropriate value at which to evaluate the selected function.

Compute. Click to compute the selected function value for the specified distribution.

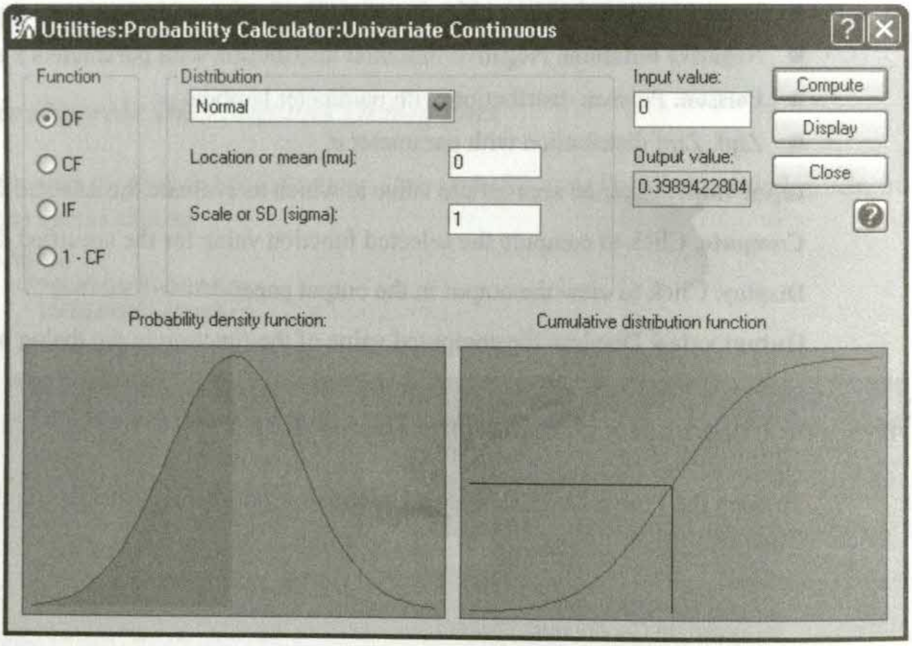
Display. Click to view the output in the output pane.

Output value. Displays the computed value of the function in the dialog box.

Univariate Continuous Distributions Dialog Box

To open the Probability Calculator: Univariate Continuous Distributions dialog box, from the menus choose:

Utilities
  Probability Calculator
    Univariate Continuous...



Function. You need to choose one of the following options:

- **DF.** Probability density function
- **CF.** Cumulative distribution function
- **IF.** Inverse cumulative distribution function
- **1-CF.** Upper-tail probability

Distribution. From the drop-down list, select a distribution, and specify its parameters. The following choices are available:

- **Beta.** Beta distribution with parameters *shape1* and *shape2*.
- **Cauchy.** Cauchy distribution with parameters *location* and *scale*.
- **Chi-square.** Chi-square distribution with parameter *df*.
- **Erlang.** Erlang distribution with parameters *shape* and *scale*.
- **Double exponential**. Double exponential distribution with parameters *location* and *scale*.
- **Exponential.** Exponential distribution with parameters *location* and *scale*.
- **F.** F distribution with parameters *df1* and *df2*.

- **Gamma.** Gamma distribution with parameters *shape* and *scale*.
- **Gompertz.** Gompertz distribution with parameters *b* and *c*.
- **Gumbel.** Gumbel distribution with parameters *location* and *scale*.
- **Inverse Gaussian.** Inverse Gaussian distribution with parameters *location* and *scale*.
- **Logistic.** Logistic distribution with parameters *location* and *scale*.
- **Logit normal.** Logit normal distribution with parameters *location* and *scale*.
- **Loglogistic.** Loglogistic distribution with parameters log of *scale* and *shape*.
- **Lognormal.** Lognormal distribution with parameters *location* and *scale*.
- **Normal.** Normal distribution with parameters *location* and *scale*.
- **Non-central chi-square.** Non-central chi-square distribution with parameters *df* and *non-centrality*.
- **Non-central F.** Non-central F distribution with parameters *df1*, *df2* and *non-centrality*.
- **Non-central t.** Non-central t distribution with parameters *df* and *non-centrality*.
- **Pareto.** Pareto distribution with parameters *threshold* and *shape*.
- **Rayleigh.** Rayleigh distribution with parameter *scale*.
- **Smallest extreme value.** Smallest extreme value distribution with parameters *location* and *scale*.
- **Studentized maximum modulus.** Studentized maximum modulus distribution with parameters *k* and *df*.
- **Studentized range.** Studentized range distribution with parameters *k* and *df*.
- **t.** t distribution with parameter *df*.
- **Triangular.** Triangular distribution with parameters *low*, *high*, and *mode*.
- **Uniform.** Uniform distribution with parameters *low* and *high*.
- **Weibull.** Weibull distribution with parameters *scale* and *shape*.

Input value. Enter an appropriate value at which to evaluate the selected function.

Compute. Click to compute the selected function value for the specified distribution.

Display. Click to view the output in the output pane.

Output value. Displays the computed value of the function in the dialog box.

Probability density function. Displays the graph of the density function.

Cumulative distribution function. Displays the graph of the distribution function.

1-Cumulative distribution function. Displays the graph of the function (1- CF).

Using Commands

Commands are not available for Probability Calculator.

Usage Considerations

Probability Calculator only uses numeric data.

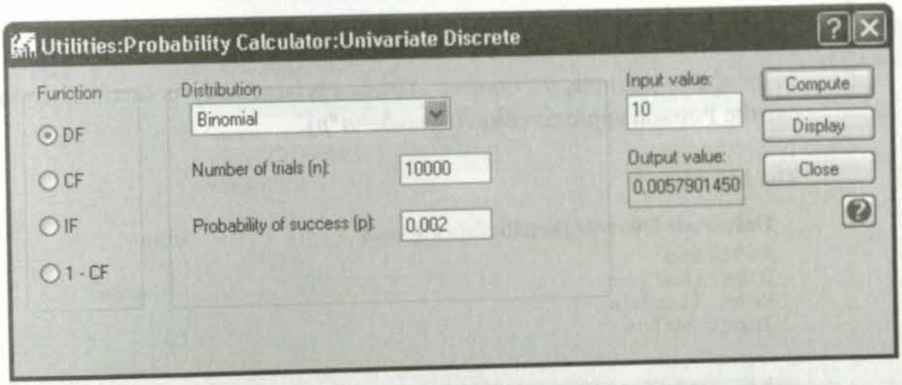
Examples

Example 1
Calculating Probability Mass Function and Cumulative Distribution Function for Discrete Distributions

The probability of a traveler getting injured in a road accident on a national highway during a certain period is 0.002. Suppose you want to compute the probability that out of 10000 travelers, exactly10 are injured during that period, assuming independence of travelers.

The input is:

Univariate Discrete Distribution Options	**Input**
Function	DF
Distribution	Binomial
Number of trials (n)	10000
Probability of success (p)	0.002
Input value	10

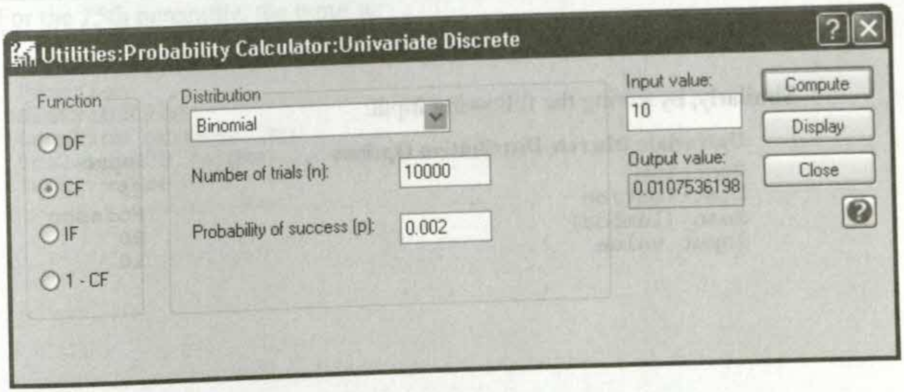


The Output value displayed is 0.0057901450.

To compute the probability that at most 10 travelers are injured:

The input is:

Univariate Discrete Distribution Options	**Input**
`Function`	`CF`
`Distribution`	`Binomial`
`Number of trials (n)`	`10000`
`Probability of success (p)`	`0.002`
`Input value`	`10`



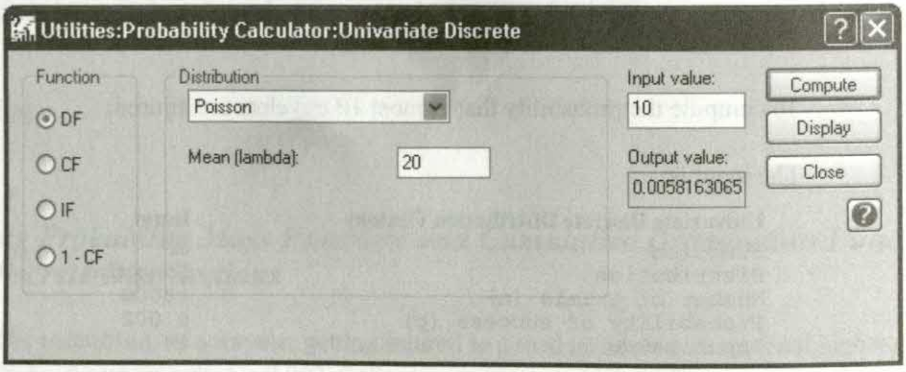
The Output value displayed is 0.0107536198.

Poisson Approximation

In the above example, the number of trials *n* is large and *p* is small. Therefore, you can use the Poisson approximation (*lambda=n*p*).

The input is:

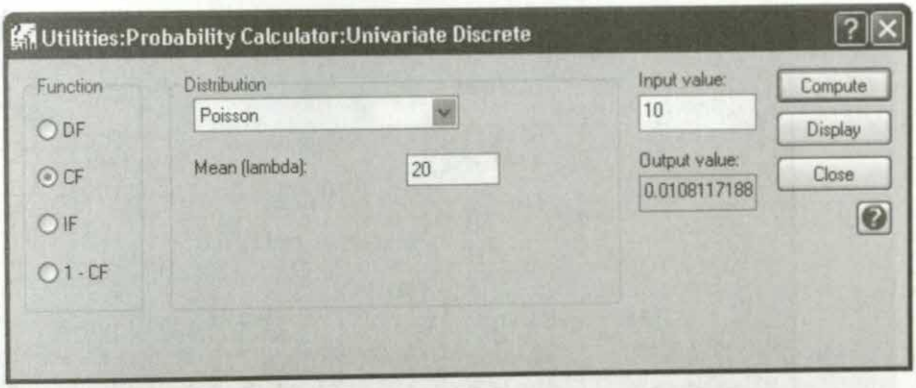
Univariate Discrete Distribution Options	Input
Function	DF
Distribution	Poisson
Mean (lambda)	20
Input value	10



The Output value displayed is 0.0058163065.

Similarly, by giving the following input:

Univariate Discrete Distribution Options	Input
Function	CF
Distribution	Poisson
Mean (lambda)	20
Input value	10



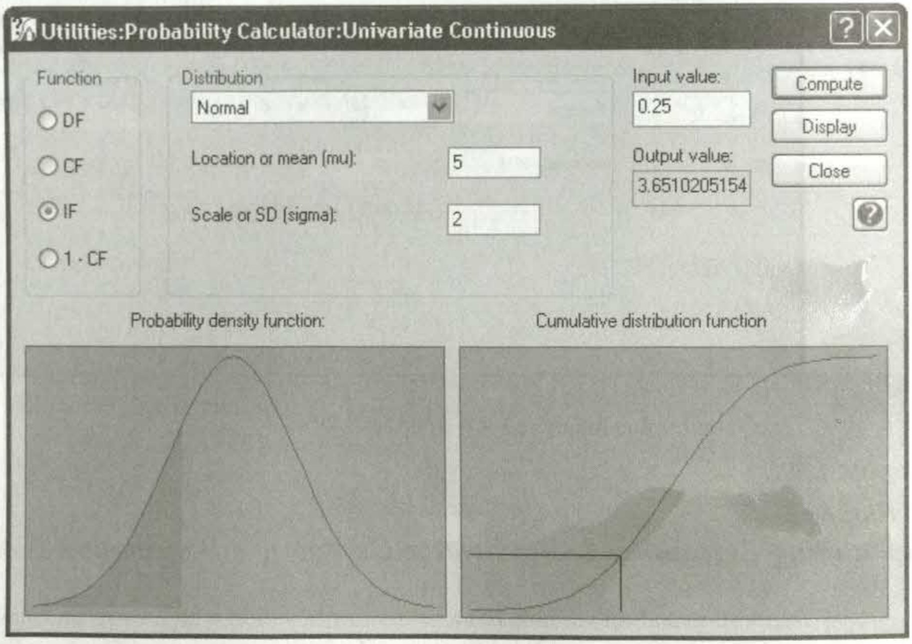
The Output value displayed is 0.0108117188.

Example 2
Calculating Percentiles Using Inverse Cumulative Distribution Function

Suppose a random variable has a normal distribution with mean 5 and standard deviation 2. To compute the 25th, 50th, and 75th percentiles of the normal distribution, you can use the IF function:

For the 25th percentile, the input is:

Univariate Continuous Distribution Options	**Input**
`Function`	`IF`
`Distribution`	`Normal`
`Location or mean (mu)`	`5`
`Scale or SD (sigma)`	`2`
`Input value`	`0.25`



The Output value displayed is 3.6510205154.

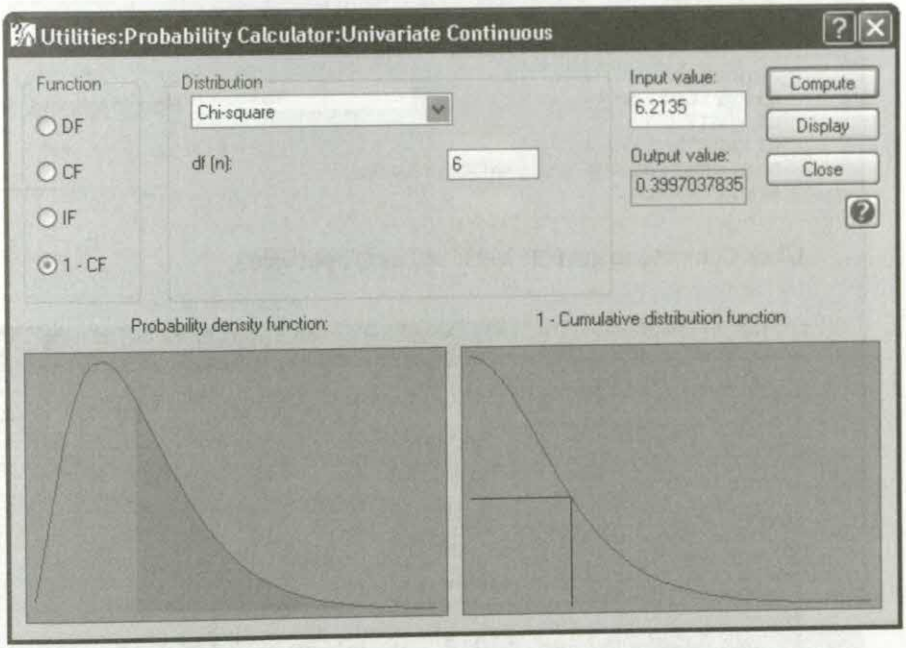
Similarly, for the 50th and 75th percentiles, the outputs are respectively 5.000000000 and 6.3489794846.

Example 3
Computation of p-value Using 1-CF Function

Suppose the observed value of a Chi-square statistic based on 6 degrees of freedom (*df*) for testing the independence of two categorical variables is 6.2135. To compute the *p-value*, you can use the 1-CF function.

The input is:

Univariate Continuous Distribution Options	**Input**
`Function`	`1-CF`
`Distribution`	`Chi-square`
`df(n)`	`6`
`Input value`	`6.2135`



The Output value displayed is 0.3997037835.

Example 4
Confidence Interval for Non-Centrality Parameter in One-Way Balanced Fixed Effect ANOVA

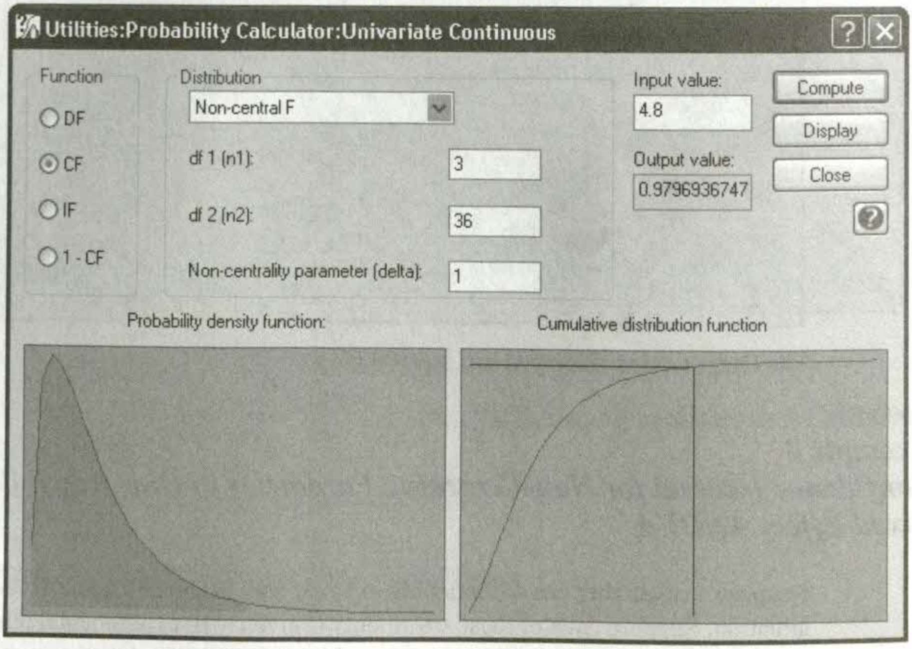
Suppose a researcher has 4 treatments in a one-way balanced fixed effect ANOVA situation. Suppose each of these treatments is given to10 experimental units, and the *F-ratio* that results is 4.8, with a *p-value* of 0.0065. The null hypothesis of equality of means is rejected at 5% level. If, further, you want to compute a 95% confidence interval for the non-centrality parameter δ of the non-central F under the alternative hypothesis, then you can use the probability calculator as described below.

The cumulative distribution function (usually denoted by CDF in the literature, but denoted by *CF* here) of a non-central F distribution is a strictly decreasing function of the non-centrality parameter δ, for a fixed *x*, *df1* and *df2*. Given this, we can find a 95% confidence interval for the non-centrality parameter δ. Following Steiger (2004), we have to find values of the non-centrality parameter for which the *CF* values are 0.975 and 0.025.

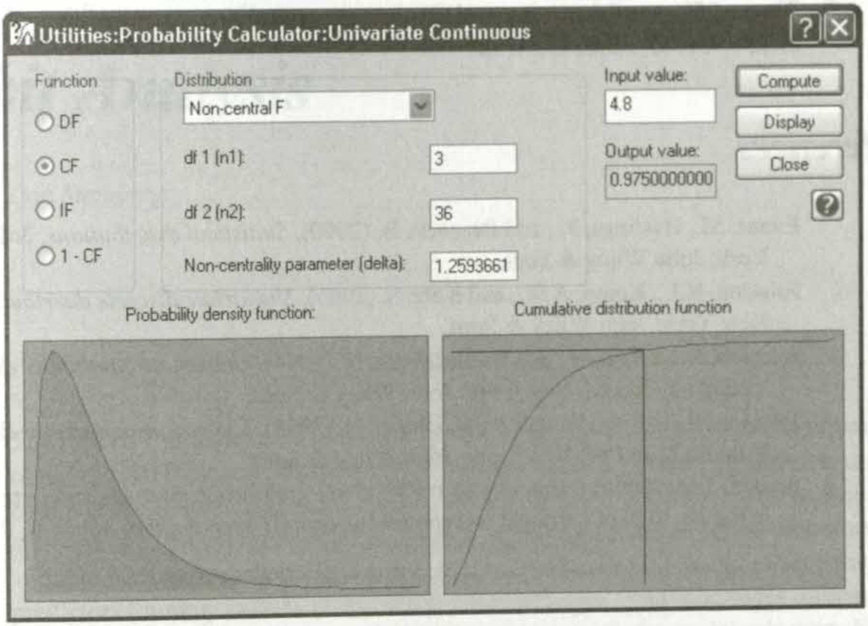
The input is:

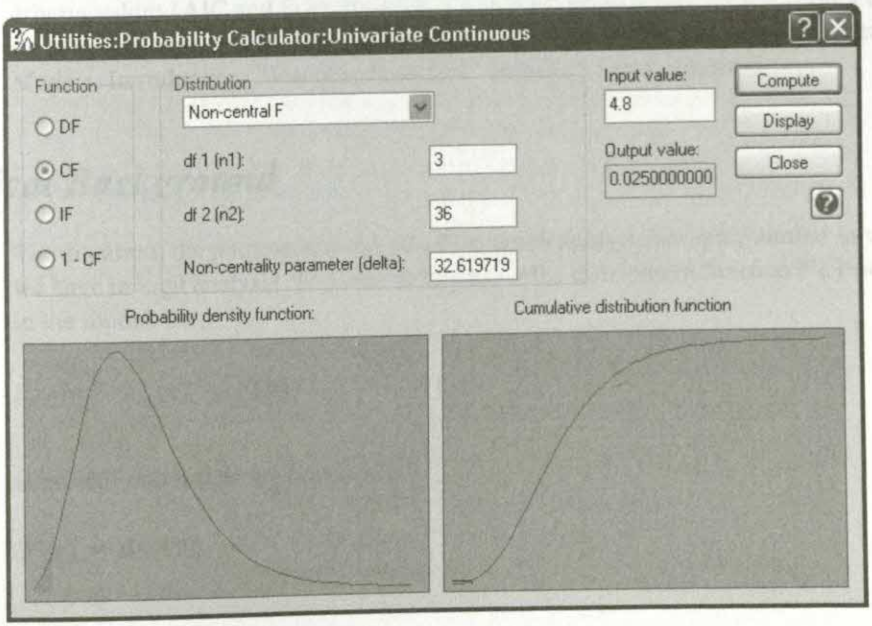
Univariate Continuous Distribution Options	**Input**
Function	CF
Distribution	Non-central F
df (n1)	3
df(n2)	36
Non-centrality parameter (delta)	1
Input value	4.8

Click Compute to get 0.9796936747 as Output value.



Now, by incrementing (decrementing) the value of δ by a suitable number (say 0.001), you will observe a decreasing (increasing) sequence of *CF* values. Finally you will get a *CF* value of 0.975 for the *Non-centrality parameter* 1.259366185, and 0.025 for the *Non-centrality parameter* 32.61971977.


Utilities:Probability Calculator:Univariate Continuous
Function
DF
CF
IF
1 - CF
Distribution
Non-central F
df 1 (n1):
3
df 2 (n2):
36
Non-centrality parameter (delta):
1.2593661
Input value:
4.8
Output value:
0.9750000000
Compute
Display
Close
Probability density function:
Cumulative distribution function


Utilities:Probability Calculator:Univariate Continuous
Function
DF
CF
IF
1 - CF
Distribution
Non-central F
df 1 (n1):
3
df 2 (n2):
36
Non-centrality parameter (delta):
32.619719
Input value:
4.8
Output value:
0.0250000000
Compute
Display
Close
Probability density function:
Cumulative distribution function

Thus a 95% confidence interval for the non-centrality parameter δ is (1.259366185, 32.61971977).

References

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical distributions*, 3rd ed. New York: John Wiley & Sons.

Johnson, N.L., Kemp, A.W., and Kotz, S. (2005). *Univariate discrete distributions*, 3rd ed. New York: John Wiley & Sons.

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1994). *Univariate continuous distributions*. Volume 1, 2nd ed. New York: John Wiley & Sons.

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1995). *Univariate continuous distributions*. Volume 2, 2nd ed. New York: John Wiley & Sons.

Steiger, J. H. (2004). Beyond F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9:2, 164-182.

Chapter

4

Probit Analysis

Dan Steinberg

The PROBIT module calculates the maximum likelihood estimates of the parameters of the PROBIT general linear model. A modified Gauss-Newton algorithm is used to compute the estimates. Conventionally, the dependent variable is coded as 0 or 1, although the PROBIT module will automatically recode values of the dependent variable because it assumes that it is categorical. Models may include categorical predictors (dummy coded), as well as interaction terms. Output includes information criteria values (AIC and Schwarz's BIC) which are tools for model selection. For more information on AIC and Schwarz's BIC in SYSTAT refer to Chapter Linear Models: Introduction: "Variable Selection" on page 15 in *Statistics II*.

Statistical Background

The situation, the purpose and the model in probit analysis are quite similar to what we have in logit analysis. The only difference is the distribution function $F(.)$ we use in the model:

$$Prob(Y = 1|x) = F(x)$$

is normal rather than logistic, and so

$$F(x) = \Phi(\boldsymbol{Xb})$$

where Φ is the cumulative normal distribution and **b** a vector of unknown regression coefficients.

The purpose of PROBIT analysis is to produce an estimate of the probability that the value of the dependent variable equals 1 for any set of independent variable values, and to identify those independent variables that are significant predictors of the outcome. The estimated coefficients, **b,** are assumed to generate a predicted z score, $\mathbf{Xb}$.

Interpreting the Results

The simplest interpretation of the output is obtained by noting the significant variables and identifying them as useful predictors of the dependent variable. More sophisticated interpretations require scaling the coefficients into derivatives. While the predicted effect of an independent variable on the z score is linear, the effect on the probability that the dependent variable equals 1 is nonlinear. The derivative of this probability with respect to the i'th independent variable is given by

$$b_i f(Xb)$$

where $f(\mathbf{Xb})$ is the normal density evaluated at the predicted z score.

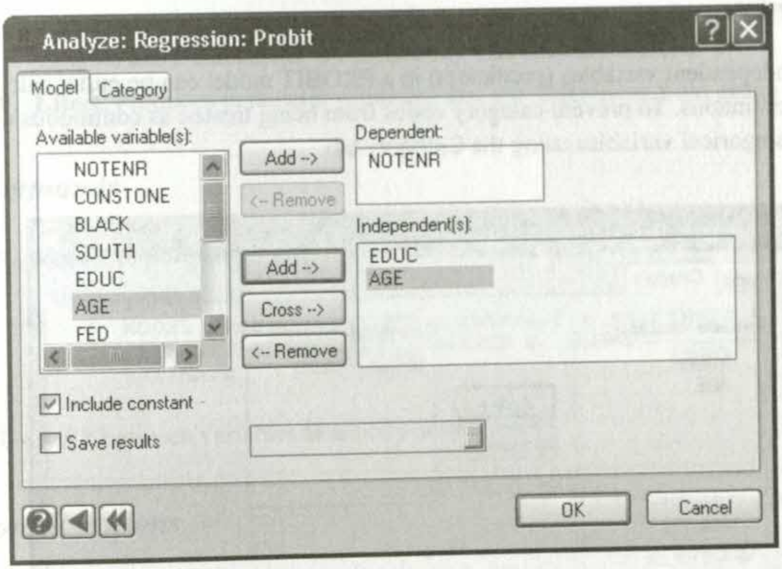
This derivative will differ for each observation in the data set. The correct way to estimate this derivative for the sample is to evaluate it for each observation and then average all the observations. A good approximation can be obtained by evaluating the s score for the mean set of X's and using the above scaling formula, or using the normal density evaluated at the x score, which would split the sample to match the observed split.

Probit Analysis in SYSTAT

Probit Regression Dialog Box

To open the Probit Regression dialog box, from the menus choose:

Analyze
  Regression
    Probit...



Dependent. Select the variable you want to examine. The dependent variable should be a binary numeric variable. Normally, the dependent variable is coded so that the larger value denotes the reference group.

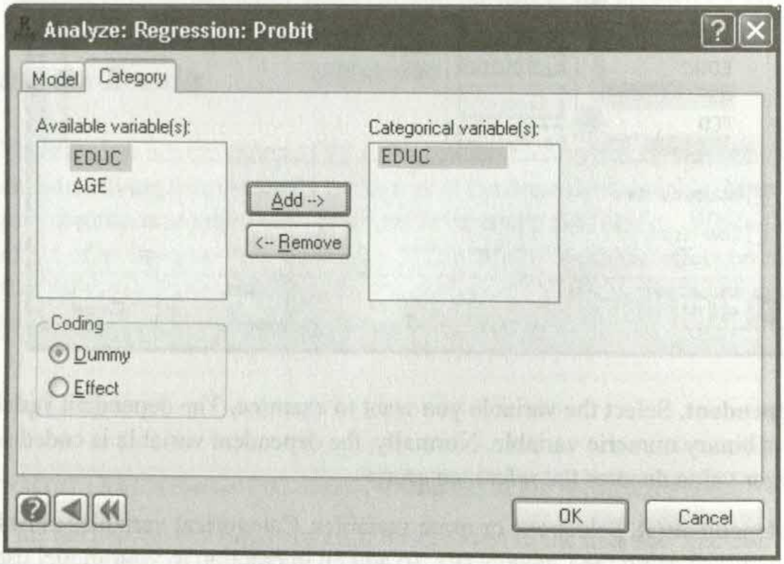
Independent(s). Select one or more variables. Categorical variables must be designated using the Category tab. To add an interaction to your model use the Cross button. For example, add *income* to the Independent list and then use the Cross button to add *education*, which will look like *income*education*. A variable with a positive correlation with the dependent variable should have a positive coefficient when fitted alone. To reverse the direction of this coding, use ORDER command with a descending sort for the dependent variable.

Include constant. Includes the constant in the regression equation. Deselect this option to remove the constant. You rarely want to remove the constant, and you should be familiar with no-constant regression terminology before considering it.

Save results. Saves statistics to a specified file.

Categories

Independent variables (predictors) in a PROBIT model can be either categorical or continuous. To prevent category codes from being treated as continuous data, specify categorical variables using the Category tab.



Available variable(s). All independent variables selected for the model appear in the variable list.

Categorical variable(s). You want to categorize an independent variable when it has several categories such as education levels, which could be divided into the following categories: less than high school, some high school, finished high school, some college, finished bachelor's degree, finished master's degree, and finished doctorate. On the other hand, a variable such as age in years would not be categorical unless age were broken up into categories such as under 21, 21–65, and over 65.

Coding. You must indicate the coding method to apply to categorical variables. The two available options include:

- **Dummy.** Produces dummy codes for the design variables instead of effect codes. Coding of dummy variables is the classic analysis of variance parameterization, in

which the sum of effects estimated for a classifying variable is 0. If your categorical variable has k categories, $k-1$ dummy variables are created.

- **Effect.** Produces parameter estimates that are differences from group means.

Using Commands

Select a data file using USE *filename* and continue with:

```
PROBIT
   MODEL yvar = CONSTANT + xvarlist + xvar*xvar +...
   CATEGORY grpvarlist / EFFECT or DUMMY
   ESTIMATE
```

Use an * between variables to specify interactions.

Usage Considerations

Types of data. PROBIT uses rectangular data only.

Print options. Output is standard for all PLENGTH options.

Quick Graphs. PROBIT produces no Quick Graphs.

Saving files. In the PROBIT model, the predicted value of the dependent variable is a normal *z* score. If you want to save this variable, you can issue a SAVE command. This will produce a SYSTAT system file with the variable's *z* score, the predicted *z* score from the last model estimated, and *MILLS*, the hazard function evaluated at the predicted *z* score. By using the cumulative normal probability function, you can convert the *z* score into a predicted probability. The *MILLS* variable is often used as a selectivity bias correction variable in regression models with nonrandom sampling. For further details, see the references at the end of this chapter.

Additional variables saved are *SEZSCORE* (standard errors), *PROB* (corresponding probability), *DENSITY* (associated density value), and confidence intervals for the parameters.

BY groups. PROBIT analyzes data by groups. Your file need not be sorted on the BY variable(s).

Case frequencies. FREQ <*variable*> increases the number of cases by the FREQ variable. This feature does not use extra memory.

Case weights. WEIGHT is not available in PROBIT.

Examples

Example 1
Probit Analysis (Simple Model)

This example shows a simple linear PROBIT model. The data, which have been extracted from the National Longitudinal Survey of Young Men, 1979, includes school enrollment status (*NOTENR* = 1 if not enrolled), age (*AGE*), highest completed grade (*EDUC*), mother's education (*MED*), an index of reading material available in the home (*CULTURE*), and an IQ score (*IQ*) for 200 individuals.

The input is:

```
PROBIT
   USE NLS
   MODEL NOTENR = CONSTANT + EDUC + AGE
   ESTIMATE
```

The output is:

```
SYSTAT Rectangular file contains variables:

  NOTENR   CONSTONE   BLACK    SOUTH   EDUC   AGE
  FED      MED        CULTURE  NSIBS   LW     IQ
  FOMY

Categorical values encountered during processing are

  Variables          :      Levels
  -------------------+-----------------
  NOTENR (2 levels)  :  0.000     1.000

Binary Probit Analysis

Dependent Variable                            : NOTENR
Input Records                                 : 200
Records kept for Analysis                     : 200

Convergence achieved after 4 iterations.

Relative Tolerance                            :   0.000
Number of Observations                        : 200.000
Number with NOTENR = 0 (Non-response)         :  28.000
Number with NOTENR = 1 (Response)             : 172.000

Results of Estimation

Log-Likelihood                                : -75.240
```

```
Information Criteria

AIC            | 156.480
Schwarz's BIC  | 166.375

Parameter Estimates

Parameter    | Estimate   Standard Error        t    p-value
-------------+----------------------------------------------
1 CONSTANT   |   2.187            1.148     1.905     0.057
2 EDUC       |  -0.161            0.051    -3.185     0.001
3 AGE        |   0.048            0.040     1.184     0.236

-2 * Log-Likelihood Ratio : 11.505
df                        : 2
p-value                   : 0.003

Covariance Matrix

  |       1        2        3
--+-------------------------
1 |   1.318
2 |  -0.027    0.003
3 |  -0.035    0.000    0.002
```

PROBIT always reports the number of cases processed and the means of the dependent variable for each of the two subgroups defined by the value of the dependent variable. If all observations have the same value of the dependent variable, PROBIT will return an error message indicating that the model cannot be estimated.

Before printing the coefficient estimates, PROBIT reports whether it has achieved convergence, the value of the likelihood function, the percentage change in the likelihood achieved in the last iteration, the convergence criterion, the number of iterations required, the size of the two subsamples of the data, and the likelihood ratio chi-square test of the null hypothesis that all coefficients except the constant are equal to 0. If there isn't a constant specified on the model statement, this last statistic will not be computed. Next, the coefficient estimates, standard errors, and t statistics are presented. The coefficients, analogous to regression coefficients, represent the change in a score that is predicted by a unit change in the independent variable. Finally, the variance-covariance matrix of the coefficient estimates is printed. This matrix, analogous to the inverse $(X'X)$ of a linear regression model, can be used to conduct hypothesis tests.

Example 2
Probit Analysis with Interactions

This example adds an interaction term to the simple model from the other example. While doing this, it is useful to standardize variables in product terms so that they do not "soak up" the variance from the main effects simply because they become highly correlated due to scale effects. You can compare the results with and without standardization.

The input is:

```
PROBIT
   USE NLS
   STANDARDIZE EDUC AGE
   MODEL NOTENR= CONSTANT + EDUC + AGE + EDUC*AGE
   ESTIMATE
```

The output is:

```
SYSTAT Rectangular file contains variables:

  NOTENR   CONSTONE   BLACK     SOUTH    EDUC   AGE
  FED      MED        CULTURE   NSIBS    LW     IQ
  FOMY

Categorical values encountered during processing are

  Variables         |        Levels
  ------------------+-------------------
  NOTENR (2 levels) |  0.000      1.000

Binary Probit Analysis

Dependent Variable                             : NOTENR
Input Records                                  : 200
Records kept for Analysis                      : 200

Convergence achieved after 4 iterations.

Relative Tolerance                             :   0.000
Number of Observations                         : 200.000
Number with NOTENR = 0 (Non-response)          :  28.000
Number with NOTENR = 1 (Response)              : 172.000

Results of Estimation

Log-Likelihood                                 : -72.805

 Information Criteria

 AIC           | 153.610
 Schwarz's BIC | 166.804
```

Parameter Estimates

Parameter	Estimate	Standard Error	t	p-value
1 CONSTANT	1.176	0.125	9.381	0.000
2 EDUC	-0.479	0.137	-3.505	0.000
3 AGE	0.067	0.126	0.530	0.596
4 EDUC*AGE	0.274	0.128	2.141	0.032

```
-2 * Log-Likelihood Ratio : 16.375
df                        : 3
p-value                   : 0.001
```

Covariance Matrix

	1	2	3	4
1	0.016			
2	-0.006	0.019		
3	0.000	0.002	0.016	
4	0.002	-0.006	-0.004	0.016

Notice that the education main effect remains significant and the interaction is itself moderately significant in this expanded model.

Computation

Algorithms

PROBIT maximizes the likelihood function for the binary PROBIT model by the Newton-Raphson method.

Missing Data

Cases with missing data for any variable in the model are deleted.

References

* Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, December, 1483–1536.

* Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, (eds.) *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267-281.

* Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* 19, 716-723.

* Burnham, K.P. and Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York: Springer-Verlag.

* Finney, D. J. (1971). *Probit analysis*, 3rd ed. Cambridge: Cambridge University Press.

* Heckman, J. (1979). Sample bias as a specification error. *Econometrica*, 47, 153–162.

* McFadden, D. (1982). Qualitative response models. In W. Hildebrand (ed.), *Advances in Econometrics,* 1-25, Cambridge: Cambridge University Press.

* Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, 461-464.

(* indicates additional references.)

Chapter

5

Quality Analysis

S.K.Adhikari, B.K.Manna, and P.R.P.Pillai
(This includes a modified version of the chapter on "Quality Control Charts" by Herb Stenson in the earlier versions.)

The Quality Control (QC) module provides 11 types of Shewhart control charts (*X*-bar, variance, *s*, *R*, *X*-bar and *s*, *X*-bar and *R*, *X*, *np*, *p*, *c* and *u*), Cumulative sum (CUSUM) chart, Moving average chart, Exponentially weighted moving average (EWMA) chart, Individual and moving range (XMR) chart, Regression chart, and Hotelling's T^2 (TSQ) chart. In Shewhart, Moving average, EWMA and X-MR charts the eight different run tests proposed by Nelson (1984) are also available to test whether certain types of non-random behavior are present in the plotted points. SYSTAT uses the statistical distribution function that is appropriate for each chart while computing the control limits. However the traditional "sigma limits" approach is also available as an option. SYSTAT also provides operating characteristic (OC) and average run-length (ARL) curves for eight statistical distributions. QC also provides process capability indices and process performance indices to assess the uniformity of a process for normal as well as the following non-normal distributions: Beta, Exponential, Gamma, Inverse Gaussian, Lognormal, Rayleigh and Weibull. Plots like Histogram and Box-and-Whisker in SYSTAT also help to assess the uniformity of the process. Pareto Charts in SYSTAT can quickly and visually identify the most frequently occurring types of defects.

The result of most of the analysis can be saved to data files for further analysis.

Statistical Background

Even in a well-designed and well-run manufacturing process, no two products are completely identical and variation is inevitable. This inherent variability is due to chance causes. There may be another kind of variability, the variability due to

assignable causes, which can be identified and rectified in future. Typically, assignable causes are due to operator errors, defective raw material, etc. A process that is operating only within chance causes is said to be in statistical control while a process operating under assignable causes is said to be out of control.

Control chart is a very useful process monitoring technique when unusual sources of variability are present. It was developed in 1920 by Dr.Walter A. Shewhart of the Bell Telephone Laboratories. Essentially control chart is a graphical display of the quality characteristic that has been measured. The chart shows a center line, an upper control limit and a lower control limit. The usual interpretation of the control limit is that as long as the plotted point falls within the control limit, the process is considered in control and no action is required. However, a point that plots outside the control limits is interpreted as evidence that the process is out of control. In such case investigation and corrective actions are required to find and eliminate the assignable causes.

Control charts can also be used to estimate the inherent variability of a process. A comparison of inherent variability in a process with the specifications or requirements for the product leads to the concept of process capability analysis. A simple way to express the process capability is through the process capability ratio. It is a measure of the ability of the process to manufacture product that meets the specification.

Quality Analysis in SYSTAT

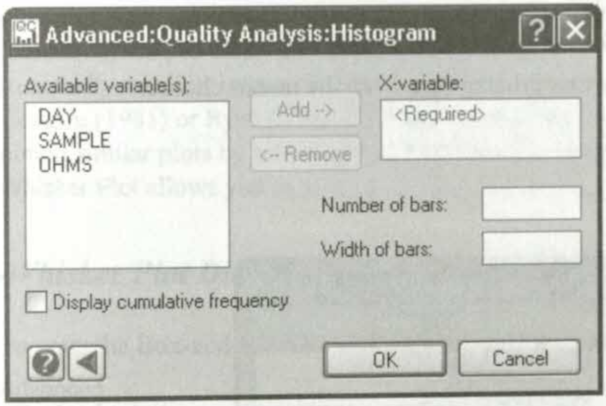
Histogram

Histograms are the graphical representation of the sample density of a continuous variable with a series of vertical bars. Here SYSTAT provides a simpler version of the Histogram in the Graph menu.

Quality Analysis: Histogram Dialog Box

To open the Histogram dialog box, from the menus choose:

Advanced
 Quality Analysis
  Histogram...



X-variable. Select an *x* variable for which you want to draw the histogram.

Number of bars. You can specify the number of bars (intervals) displayed. The size of each interval depends on the range of the data. The maximum number of bars that SYSTAT provides in the output is 200.

Width of bars. You can specify the width of each interval. The actual number of bars depends on the range of the data. If you also specify the number of bars, the width option takes precedence.

Display cumulative frequency. Each bar's area is the sum of the preceding bar's area and its own incremental area. This makes a cumulative histogram correspond to a cumulative frequency distribution or for continuous data, a distribution function.

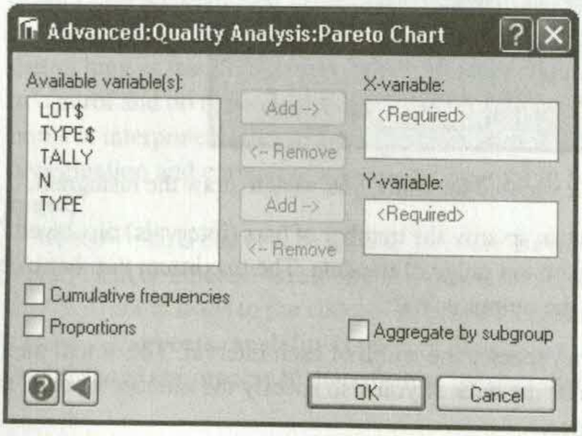
Pareto Charts

Pareto produces a chart showing frequencies of occurrence sorted in descending order, as a function of an *x* variable that identifies each sample. The *y* variable must contain zeros and ones indicating individual instances of an event, unless the Aggregate by subgroup option is selected.

Pareto Chart Dialog Box

To open the Pareto Chart dialog box, from the menus choose:

Advanced
 Quality Analysis
  Pareto Chart...



X-variable. Select the variable you want to chose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Cumulative frequencies. Displays cumulative frequencies rather than frequencies by sample.

Proportions. Displays relative frequencies rather than frequencies. Select this option together with cumulative frequencies to get cumulative proportions.

Aggregate by subgroup. Indicates that the *y* variable contains the number of instances of an event already aggregated by sample.

Box-and-Whisker Plots

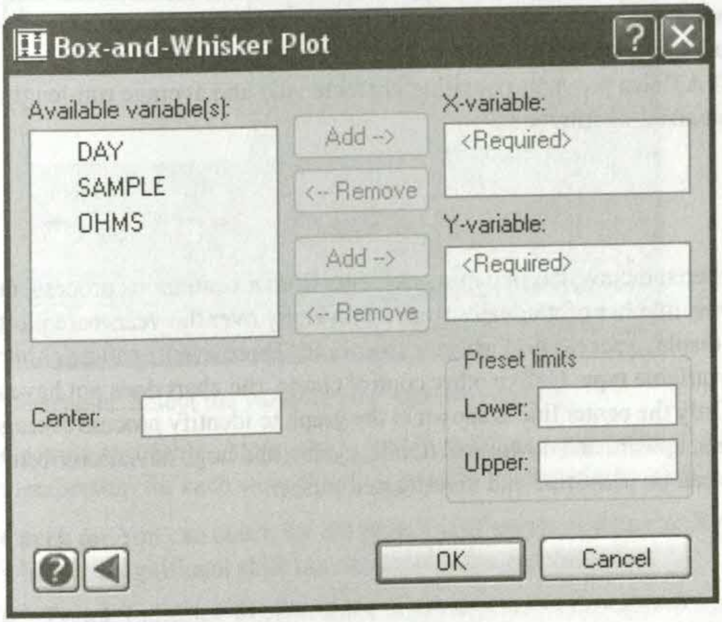
Box plots or "box-and-whisker" plots provide a convenient way to inspect the characteristics of entire statistical distributions visually. The charts show a series of box plots for the distribution of a *y* variable as a function of an *x* variable that identifies

individual samples.The chart enables simultaneous control of central tendency (medians), sample variability (inter-quartile range), and the detection of outliers (unusually small or large values for individual cases in a sample). See Velleman and Hoaglin (1981) or Ryan (2000) for basic information on box plots. Note that you can obtain similar plots by selecting Box Plot from the Graph menu; however, Box-and-Whisker Plot allows you to display a center line and control limits on the chart.

Box-and-Whisker Plot Dialog Box

To open the Box-and-Whisker Plot dialog box, from the menus choose:

Advanced
  Quality Analysis
    Box-and-Whisker Plot...



X-variable. Select the variable you want to chose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Center. Displays a center line at the specified value. If you do not specify a value, no center line is displayed.

Preset limits. Specify Lower and Upper values for control limits on the chart. With commands, you can use LIMITS to specify up to12 values.

Control Charts

SYSTAT provides eleven types of Shewhart charts (*X*-bar, variance, *s*, *R*, *X*-bar and *s*, *X*-bar and *R*, *X*, *np*, *p*, *c*, and *u*) as well as Run, Cumulative Sum, Moving Average, Exponentially Weighted Moving Average, X-MR, Regression, and TSQ (Hotelling's T^2) charts.

Before computers were widely available, it was necessary to use the normal distribution as an approximation to many of the distributions needed for Shewhart charts. This situation no longer holds and SYSTAT uses the statistical distribution that is appropriate for each chart. However, the more traditional "sigma limits" approach is also available as an option.

SYSTAT also provides operating characteristic and average run-length curves for eight statistical distributions.

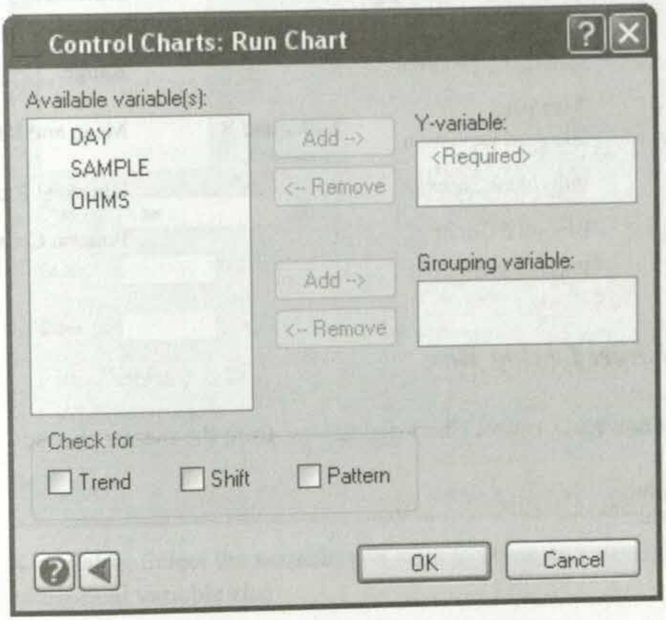
Run Charts

In run charts, the events (like measurements from a continuous process, number of defectives, number of graduates from a university over the years) are plotted against time to display process performance over time. These events can be either attribute type or variable type. Unlike other control charts, run chart does not have any control limits. Only the center line is shown in the graph to identify process average. By using run charts, upward and downward trends, cycles, and large deviations from the process average can be identified and investigated further.

Run Chart Dialog Box

To open the Run Chart dialog box, from the menus choose:

Advanced
Quality Analysis
Control Charts
Run Chart...



Y-variable. Select the variable you want to examine.

Grouping variable. A grouping variable contains a value that identifies group membership for each case. This is generally a categorical variable.

Check for. You can check for the presence of trends or patterns. You can also check whether a significant shift has occurred in the process mean.

- **Trend.** Indicates whether there is any upward or downward trend or not.
- **Shift.** Indicates whether the process has shifted upward or downward or not.
- **Pattern.** Indicates whether any specific pattern exists or not.

Shewhart Control Charts

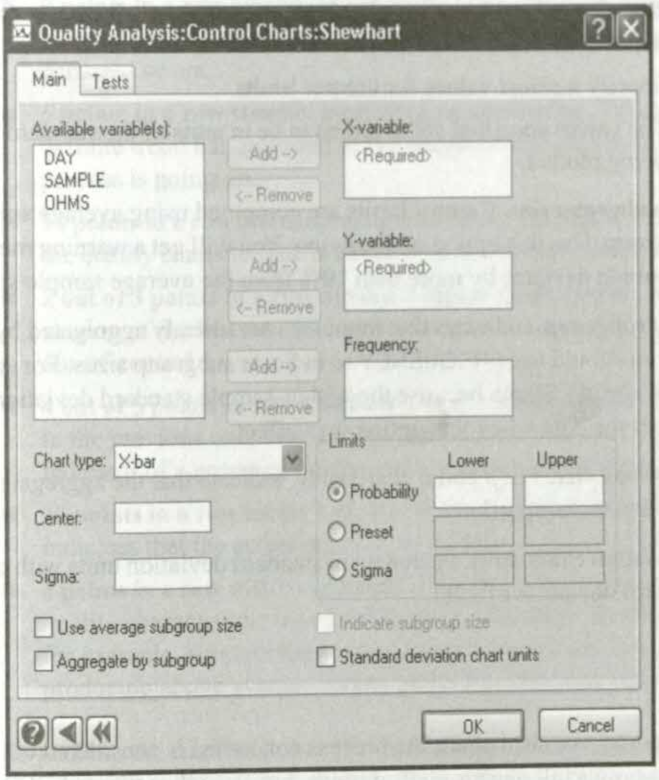
SYSTAT produces 11 types of Shewhart charts. Each chart shows a summary statistic, such as the sample mean, plotted for each sample, center line and control limits for the statistic plotted, and also the results of run tests. The statistic plotted depends on the chart type, as indicated below:

Type	Statistics	Type	Statistics
X-Bar	Mean	Variance	Variance
s	Standard Deviation	R	Range
X-Bar and s	Mean and Standard Deviation	X-Bar and R	Mean and Range
X	Individual Cases	p	Binomial Proportion
np	Binomial Count	c	Poisson Count
u	Poisson Rate		

Shewhart Control Chart Dialog Box

To open the Shewhart Control Chart dialog box, from the menus choose:

Advanced
 Quality Analysis
  Control Charts
   Shewhart...



X-variable. Select the variable you want to chose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Frequency. Select the FREQUENCY variable to indicate subgroup sizes if you check the Aggregate by subgroup option.

Chart type. Select the type of chart you want to choose from the list. The available chart types are listed above along with the statistic plotted.

Center. An *a priori* value to use for the center line of the chart. If no value is specified, the center line is computed from the data.

Sigma. An *a priori* value for the population within-sample standard deviation. If no value is specified, sigma is computed from the data.

Limits. You can choose from Probability, Preset, or Sigma limits.

- **Probability.** Enter the proportion of the sample statistic values expected to be below each control limit.
- **Preset.** Specify a *priori* values for control limits.
- **Sigma.** The values specified are assumed to be in units of the standard error of the statistic being plotted.

Use average subgroup size. Control limits are computed using average subgroup size for each subgroup. Use this option with caution. You will get a warning message if the size of any sample deviates by more than 10% from the average sample size.

Aggregate by subgroup. Indicates that input data are already aggregated by subgroup. In this case you should use FREQUENCY to indicate subgroup sizes. For some charts, you must also specify Sigma because the within-sample standard deviation cannot be computed from the data when this option is in effect.

Indicate subgroup size. For *p* and *u* charts only, indicate that the aggregate is the total count rather than the proportion.

Standard deviation chart units. Print data in standard deviation units with center set to 0 and a standard deviation of 1.00.

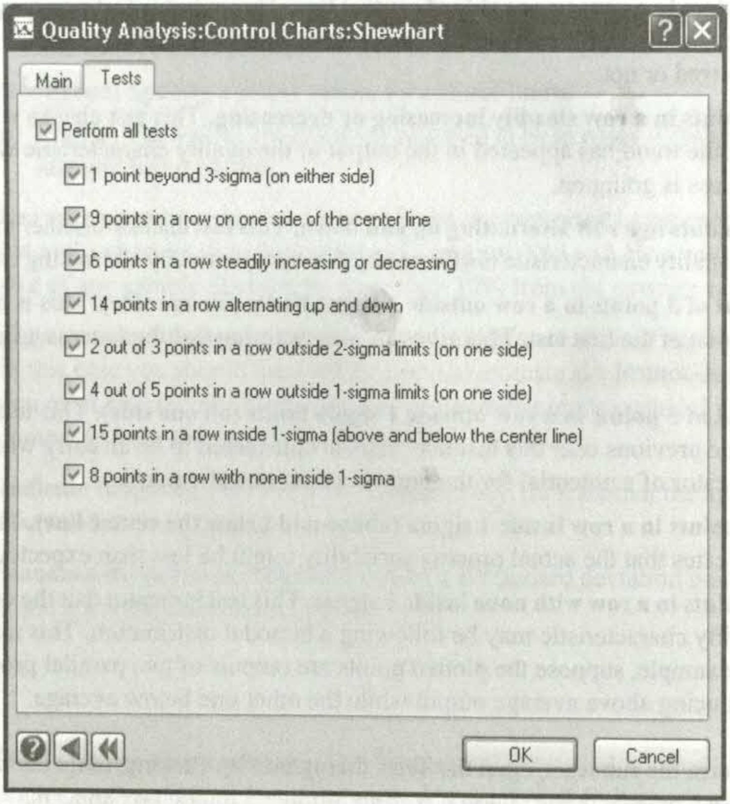
Tests

While interpreting a control chart, the process concerned is considered out of control, not only when points fall outside control limits, but also when points within the control limits do not exhibit a random behavior. The run tests (or simply, tests) proposed by Nelson (1984) in quality control charts are meant to test certain types of non-random behavior. Any pattern in the points would indicate non-random behavior and the run tests check for certain (in fact, eight in number) types of pattern. These are not statistical tests of hypotheses, but detection of certain patterns exhibited by the plotted points. It should be borne in mind, however, that the run test violation does not always indicate that the process shift has occurred exactly at that point. For example, when run test 2 (nine points on one side of 1-α limits) is violated, the shift may have occurred nine points (or more, or less) prior to the point which violated the run test. SYSTAT provides the following run tests in Shewhart, Moving Average, EWMA, and X-MR charts.

- **1 point beyond 3 sigma (on either side).** This is the most common run test. This test actually indicates that the process may be going out-of-control.

- **9 points in a row on one side of central line.** Through this test one can check whether any significant shift in the process (from the existing process mean) has occurred or not.
- **6 points in a row steadily increasing or decreasing.** This test checks whether any definite trend has appeared in the output of the quality characteristic while the process is going on.
- **14 points in a row alternating up and down.** This test checks whether the output of the quality characteristic is affected by two symmetrically alternating causes or not.
- **2 out of 3 points in a row outside 2 sigma limits (on one side).** This is a modified version of the first test. This gives an early warning that the process might be going out-of-control.
- **4 out of 5 points in a row outside 1 sigma limits (on one side).** This test is similar to the previous one; this test may also be considered to be an early warning indicator of a potential for the process going out-of-control.
- **15 points in a row inside 1 sigma (above and below the center line).** This test indicates that the actual process variability might be less than expected.
- **8 points in a row with none inside 1 sigma.** This test indicates that the output of the quality characteristic may be following a bimodal distribution. This may happen, for example, suppose the plotted points are outputs of two parallel processes, one producing above average output while the other one below average.

To perform the run tests, open the Tests dialog box by clicking Tests tab in the Shewhart Control Charts dialog box. The following dialog box appears:



Raw Data versus Aggregated Data

For each of the chart types, you may use either raw data or aggregated data. By default, SYSTAT assumes that input data are raw data consisting of individual cases for each sample. Use the Aggregate by subgroup option if you have aggregate data consisting of the statistic that is to be plotted for each sample.

For example, an X-bar chart accepts the sample mean as input for each sample. The easiest way to think about the meaning of aggregated data is to refer to the table of types. Aggregated data consists of the statistic that is plotted.

Every Shewhart chart, and most other quality control charts, accepts either a category or a numerical variable as the designator of subgroups on the x axis. If you use the CATEGORY designation, then values of x appear in the same order as they

appear in your file. If x is a numerical variable, then the x values appear as scale values on the x axis.

SYSTAT searches out each instance of a sample identifier. That is, the file does not have to be sorted by the variable that identifies the sample. This "pooling principle" applies to raw input data for any Shewhart chart. However, it applies to aggregated input data for only some of the charts.

Control Limits

Default control limits are computed using the statistical distribution appropriate to the chart, as opposed to using the normal approximation described in many quality control textbooks. Thus, the control limits are probability limits as opposed to sigma limits. Probability limits are based on the desired probability of a Type I error (alpha), using the appropriate distribution, instead of being computed from some multiple of the estimated standard error. See Montgomery (2001) and Ryan (2000).

SYSTAT may give you results that are somewhat different from other programs or from manual calculations from the traditional formulas, but SYSTAT's results are more accurate. SYSTAT allows you to specify your own limit values if you wish. You can also specify sigma limits. If you choose this method, beware of inherently skewed distributions, such as the distribution of the sample variance or a binomial distribution with a small p value, you may get misleading control limits. With commands, you can draw up to 12 separate control limits on most types of chart; use the PLIMITS, LIMITS, or SLIMITS options.

Discrete Control Limits

The usual interpretation of a control limit is that a data point that falls right at the limit is considered in control. Only points that fall outside the limit are out of control. This strictness of definition makes no difference when dealing with a statistic based on a continuous characteristic. However, the precise definition becomes important with discrete distributions such as the binomial, where the true value of alpha can be severely affected if one treats a sample value that is right at the limit as being out of control rather than in control. For the *np* and *c* charts, probability-based control limits are always placed at an integer value because outcomes for these charts must be integers. Only data that fall strictly outside the limits are considered out of control. The rules used by SYSTAT to set probability limits for discrete distributions follow.

For an upper probability limit, SYSTAT finds the smallest integer beyond which no more than a specified proportion of the distribution lies and uses that integer as the upper control limit. For a lower probability limit, SYSTAT finds the largest integer below which no more than a specified proportion of the distribution lies and uses that integer as the lower control limit. Notice that for small samples or badly skewed distributions, the lower limit may be 0. In the case of the binomial distribution, the upper limit may be the sample size. If both of these conditions exist, no sample can ever be out of control.

For example, this happens if you use the default probability limits for an *np* chart when the sample size is 2 and the population *p* value is 0.5. The probability of a count of exactly 2 is 0.25. Therefore, if one treats a count of 2 as out of control, the true upper-tail probability limit is 0.75. This is greater than the default probability limit, so SYSTAT sets the upper limit at exactly 2, meaning that a count of 2 or less is to be considered in control. Similarly, the probability of a count of exactly 0 is also 0.25, so the lower limit is set at 0, where a count of 0 or more is considered in control. The upper and lower limits found here thus demand that no sample be rejected as out of control. The true value of alpha is, in this case, 0.

Similar procedures apply to the *p* and *u* charts, which are also discrete because they are forms of the binomial and Poisson, respectively. The only difference is that an integer value is found for the proper binomial or Poisson and then converted to a proportion or rate.

There are, of course, other approaches to setting probability limits for discrete distributions. The appeal of SYSTAT's approach is that the actual probability of a Type I error, alpha, never exceeds the specified value, although it might be less than that value.To set the limits differently, use the Sigma or Preset limits options (SLIMITS or LIMITS). You can use SYSTAT's Probability Calculator feature to aid in finding probability limit values.

Sigma Limits

SYSTAT computes probability limits rather than sigma limits, so that the default control limits are always set with reference to alpha. The appropriate distribution is always used to set these probability limits. For all Shewhart charts, a Sigma limits (SLIMITS) option is available for computing the traditional sigma limits. Some of the examples show that sigma limits can lead to a different course of action than probability limits. We offer the following arguments if you are inclined to stick with the traditional approach and use sigma limits.

Using 3-sigma limits, or sigma limits in general, implies a normal distribution approximation to the distribution for which control limits are desired. Notice that if you are plotting an *X*-bar chart and you specify sigma limits without any values, the 3-sigma limits are produced. If you desired 2.5-sigma limits, you could, for the normal distribution, find the value of alpha that produces 2.5-sigma limits and so on for any sigma limits that you want. Therefore, for the *X*-bar chart, the two approaches produce exactly the same limits.

For other charts, consider the following argument. If the normal distribution is really a good approximation to the distribution that is required, the default alpha will produce limits very close to the 3-sigma limits obtained using the traditional approach. The better the normal approximation, the closer the values obtained by the two approaches. However, if the normal approximation is poor, you should consider the SYSTAT probability limits as the more accurate, unless you have no concern for the actual value of alpha. You may, in fact, have no such concern. If you need a particular set of control limits regardless of alpha, you should use the Preset limits (LIMITS) option to specify them.

The only case in which the arguments given above are not valid is when the true distribution of the statistic being plotted does not conform to the model being used. For example, if sample variances are computed from individual cases that came from a non-normal distribution, then the chi-squared distribution is an inappropriate model for the distribution of sample variances. In such a case, the appropriate sampling distribution is usually unknown. Then the normal approximation implied by the sigma limits approach is not necessarily any more accurate than is the probability limits approach. We simply do not know what distribution to use as a model in this case.

X-Bar Charts

An X-bar chart plots sample means of the specified *y* variable as a function of the *x* variable, a sample identifier. The default center line of an *X*-bar chart is always either the mean of all of the *y* variable values used for a chart or 0 if you request a Standard deviation chart.

The default control limit(s) are based on probability limits, sigma, and the sample size. The program finds the standard normal deviate below which a specified proportion *p* of the normal distribution lies. This value is multiplied by SIGMA/sqr(n) where n is the sample size. The center value is subtracted from this value to produce a control limit. These calculations are bypassed if you specify values for preset or sigma limits.

If the sample sizes differ, the control limit(s) are not constant for all samples, producing an unaesthetic chart. One remedy for this is the Use average subgroup size (AVGN) option; however, this option causes the subgroup size to be wrong for every sample, strictly speaking, so it should be used with caution. The Standard deviation chart units (Z) option produces the same effect without compromising the data.

SYSTAT searches out each instance of a sample identifier. That is, the file does not have to be sorted by the variable that identifies the sample. In the examples, we were able to plot by *DAY* and then by *SAMPLE* using the same file for both charts. For the *X*-bar chart, this principle also applies when you use aggregated data as input. Suppose, for example, you have sample means as cases and that each mean is identified by sample number within a day and by day. If you want to plot by day, the program pools all of the sample means for each day into a single mean. If you then plot by sample, the program pools the means for each sample number across days, thus providing a mean for each sample number regardless of the day.

Using the average range to estimate sigma. Another form of the *X*-bar chart discussed in many textbooks uses the mean sample range to estimate the standard deviation needed to compute control limits. The range is an inefficient estimator of the population standard deviation, and its distribution is difficult to deal with computationally. The usual normal approximation is poor because the sampling distribution of the range remains quite skewed even for fairly large samples. For these reasons, SYSTAT does not provide range computations directly for the *X*-bar chart.

If you are stuck with sample ranges and do not have the individual observations, you can use the following indirect method. Use the file containing the sample ranges to produce an *R* chart, using the Aggregate by subgroup (AGG) option to indicate that the data are already ranges. The first output screen for the *R* chart states a value for estimated population standard deviation. Specify this value for Sigma when producing the desired *X*-bar chart. This technique uses the actual sampling distribution of the Gaussian range to estimate the population standard deviation, as opposed to the normal approximation. However, this is still an inefficient estimate of the standard deviation.

Variance Charts

The variance chart plots sample variances as a function of a sample identifier. Each data point on a variance chart is the within-sample variance for the designated sample. These variances are the sums of squared deviations from the sample mean divided by n-1, where n is the sample size for a given sample. The default center line is the weighted average of all of these within-sample variances, weighted by n-1 for each

sample. Thus, the default center line is the square of the value reported as *Estimated Population Standard Deviation*. If you specify Sigma, its squared value is used as the center line of the chart. If you specify Center, its value is the center line of the chart, and the value of sigma used in further calculations is the square root of the center value.

The default two-tailed control limits are based on probability limits, sigma, and the sample size, *n*. To find the lower control limit, SYSTAT finds the value of a chi-square distribution, with *n*-1 degrees of freedom, below which a proportion $\alpha/2$ of the distribution lies. This value is multiplied by the square of sigma and divided by *n-1* (The upper control limit is found in an analogous manner). If only an upper limit or only a lower limit is requested, then the same calculations described above are performed, except that $\alpha/2$ is replaced with α. Notice that *n* can differ from sample to sample. If it does, the control limits will not be constant for all samples.

If you use aggregated data as input for a variance chart, you may not have more than one instance of a sample variance associated with any given sample identifier. That is, unlike the pooling of means for the *X*-bar chart, instances of multiple variances are not pooled. Why? This is because each sample variance is computed around the mean for that subgroup of data. Pooling such variances across subgroups within a sample does not give the same answer as computing the sample variance for the entire sample around a single sample mean. You get an error message if the program finds more than one instance of the same sample identifier when the input data are already aggregated as sample variances.

This is one chart where the injudicious use of sigma limits can produce very misleading results for small samples, and results will be somewhat in error even for larger samples. This is illustrated by the contradictory results produced in two examples above. The sampling distribution of the sample variance is badly skewed for small samples because it is based on the chi-square distribution. Thus, putting limits on some fixed distance on either side of the center line causes more of the distribution to be outside the limit in one tail of the distribution than in the other (the distribution is asymmetric). If you rely on 3-sigma limits to encompass nearly all of such a distribution, your results will be wrong for small samples, and somewhat wrong for even large samples. For large samples, the sampling distribution begins to approach the normal distribution in shape, but it is still skewed.

Variance Chart Options

Here are a few specific comments about options for variance charts:

Sigma and Center. You cannot use both of these options at the same time for this chart because a stated value for one necessarily implies a fixed value for the other.

Aggregate by subgroup. This option assumes that the data are aggregated as variances. You do not need to specify Sigma, as was required for *X*-bar charts, because the default sigma is a weighted average of the variances that are input for the individual samples. Use FREQUENCY to indicate sample size.

s Charts

The *s* chart plots sample standard deviations as a function of a sample identifier. The default center line for an *s* chart is computed by multiplying the estimated population standard deviation, given in the output, by a number usually called c_4 in quality control work. See Ryan (2000) or Montgomery (2001) for discussion and the definition of c_4. The value of c_4 is a function of sample size and so the values of the center line could differ if sample sizes vary. If you supply a Sigma for this chart, that value is used as the population standard deviation for calculating the default center line.

If you specify a Center value, it is used as the center line for all samples. The value you specify, along with sample size, determines the value of sigma that is used for further calculations. Sigma is computed as the value of center divided by c_4. That is why you cannot specify both Center and Sigma. As c_4 depends on sample size, be aware of the implications of specifying a center value when you have unequal sample sizes. If the sample sizes differ, specifying a constant center causes a different value of sigma to be used in the calculation of default control limits for each sample. This does not make sense for most applications, but the location, interpretation, and implications of the center line are up to you when you specify Center.

Default control limits are computed as follows: Let the sample size for a given sample be denoted by *n*. The percentage point of a chi-square distribution with $n-1$ degrees of freedom is found so that p_1 of the distribution lies below that percentage point, where p_1 is the lowest specified probability limit (PLIMIT). This percentage point is then divided by $n-1$, and the square root of this result is multiplied by the specified or default sigma to get the lower control limit. The upper limit is found in a similar manner, using the largest specified probability limit. Thus, you will find that the control limits for an *s* chart are equal to the square root of the comparable limits for a variance chart. However, the center line of the *s* chart is not simply the square root of the center line for the comparable variance chart. The rationale for this is discussed next.

The default center line is computed as though the value of estimated population standard deviation in the output is the population standard deviation. We adhere to the statistical convention of using the square root of the pooled sample variances as a default estimate of the population standard deviation, even though this is a biased estimate. This is the value listed as *Estimated Population Standard Deviation* in the output. For consistency, we do this for all Shewhart charts involving continuous variables. However, this presents a theoretical problem for computing the center line and limits for the *s* chart.

The center line of the *s* chart is an estimate of the expected value of the sampling distribution of *s*, the sample standard deviation. The usual unbiased estimate of this expected value is the mean of all the sample standard deviations (assuming equal sample sizes). This expected value is not an unbiased estimate of the population standard deviation. To be unbiased, the value must be divided by c_4 (assuming a normal population distribution). If we proceed this way to make an *s* chart, then there are two different estimates of the population sigma value: one using the average sample standard deviation divided by c_4; the other using the convention described in the previous paragraph. This issue is resolved by always treating the estimated population standard deviation as the only estimate of the population sigma.

If you use aggregated data for input to the *s* chart, you may have only one sample standard deviation associated with any given sample identifier. That is, unlike the pooling of means for the *X*-bar chart, instances of multiple standard deviations are not pooled. This is because each sample standard deviation is computed around the mean for that subgroup of data. Pooling across subgroups within a sample does not give the same answer as computing the sample standard deviation for the entire sample around a single sample mean. You get an error message if the program finds more than one instance of the same sample identifier when the input data are aggregated as sample variances.

s Chart Options

Here are a few specific comments about options for *s* charts:

Sigma and Center. You cannot use both of these options at the same time for this chart because a stated value for one necessarily implies a fixed value for the other.

Control limits. The default control limits are probability limits using the sampling distribution of the sample standard deviations with $p = 0.0027$ (0.00135 in each tail). If you specify sigma limits, they are scaled in standard deviation limits.

Aggregate by subgroup. This option signals that input data in the *y* variable are already aggregated as standard deviations, one for each sample. You do not need to specify a value for Sigma with this option for this chart type because the default sigma is a weighted average of the variances that input for the individual samples. Use FREQUENCY to indicate sample size.

R Charts

The *R* chart plots sample ranges as a function of a sample identifier. SYSTAT first calculates the expected value of the range for a standard normal distribution, given the sample size. This number is usually called the relative range. Then SYSTAT uses your specified value of sigma for the within-sample population standard deviation or uses its default, which is displayed in the output as the *Estimated Population Standard Deviation*.

If you select Aggregate by subgroup (AGG) the input file should contain the range for each sample. If you do not specify a center or sigma value, sigma is obtained by dividing the mean of all the sample ranges by the standardized expected range described above. If you specify a value for Center, sigma is obtained by dividing this value by the standardized expected range. If you specify a sigma, it is used. If you do not specify a center value, the center line is computed as sigma multiplied by the standardized expected range.

When computing probability limits as control limits, SYSTAT first finds the appropriate upper and/or lower percentage points of the distribution of the standardized normal range. By default, these percentage points are dictated by probability limits. Those values are then multiplied by sigma to obtain the control limits. If you specify preset limits, then no further computation is required. If you use sigma limits, then the corresponding sigma limits are computed and used on the chart.

The sample sizes must not vary from sample to sample for this chart. If they vary slightly, you can use Aggregate by subgroup to force SYSTAT to use the average sample size in all computations. Use this option with caution.

R Chart Options

Here are a few specific comments about options for *R* charts:

Sigma and Center. These options cannot both be used for this chart.

Control limits. The default control limits are probability limits using the sampling distribution of the range of a normal distribution for a given sample size.

Aggregate by subgroup. Indicates that input data in the *y* variable are already aggregated as ranges, one for each sample. You do not need to specify a value for Sigma (when you use this option) because the default sigma is a function of the ranges that are input for the individual samples. Use FREQUENCY to indicate sample size. Sample sizes must be the same for all samples.

X-bar and *s* and *X*-bar and *R* Charts

These options allow you to plot two charts on the same screen so that you can view them simultaneously. The upper chart on the screen is an *X*-bar chart and the lower chart is either an *s* or *R* chart. The options for these charts are similar to the options for other Shewhart charts except that Preset limits, Center, and Aggregate by subgroup cannot be used because these options apply meaningfully to only one chart at a time.

X Charts

The *X* chart is identical in form to the *X*-bar chart, except that the *X* chart is specifically designed to handle the case where only one observation is available for each sample or subgroup. It is generally known as an individual cases chart. It is included as a separate chart because *X*-bar computes its estimate of the population variance from the pooled within-group variation for each sample or subgroup. If there is only one case per subgroup, then, of course, there is no within-group variation. The population variance for the *X* chart is computed as the total variance about the mean of all data. You can think of an *X* chart as an *X*-bar chart with only one observation per subgroup. If the program finds more than one observation per subgroup, you will get an error message.

np Charts

The *np* chart plots sample counts as a function of a sample identifier, using the binomial distribution as a model (Each item is classed as either conforming or non-conforming). The default center line for an *np* chart is the total number of cases in a sample multiplied by the proportion of “successes” over all samples, however you define a success. The default probability limits are computed by attempting to find integer values of the binomial variable below which p_1 and p_2 of the distribution lay,

where p_1 and p_2 are the values in effect for probability limits (PLIMITS). However, it could be that no such integers exist for a particular p because the binomial distribution is discrete. In this case, control limits are set at integer values as close to the desired p as possible without exceeding it.

Because the limits are integers only, the value of alpha may differ from the specified value. The actual alpha is the probability of a count being strictly outside the integer control limits.

The following should be kept in mind about using Aggregate by subgroup with binomial charts. This option always produces constant control limits and a constant center line, but it can alter the data values plotted on the chart. If the average sample size is used, SYSTAT divides the total number of successes for a sample by the actual sample size for that sample to produce p, the proportion of successes, and then multiplies p by the average sample size. These values are then plotted as the data. Note that these pseudo-data points may have non-integer values. To see why this procedure is necessary, imagine a case where the number of successes for a sample is 20, but the average sample size is 18.5. The number of successes cannot exceed the number used as the sample size. SYSTAT assures that this does not happen.

np Chart Options

Here are a few specific comments about options for *np* charts:

Sigma. This is not an option for this chart.

Control limits. The default control limits are probability limits using the binomial distribution for a given sample size. You can specify your own probability limits, preset control limits, or traditional sigma limits for the chart.

Aggregate by subgroup. Input data are already aggregated as counts, one for each sample or subgroup. Use FREQUENCY to indicate sample size. Otherwise, a sample size of 1 is assumed for all samples.

p Charts

The *p* chart plots sample proportions as a function of a sample identifier, using the binomial distribution of proportions as a model. The default center line is at 0.060. Here *p* is the overall proportion of "successes" without regard to the sample, however you define a success. The default control limits are computed as they are for the *np* chart, except that the resulting limits are divided by the sample size to convert them to

proportions. Only a finite set of probability limits is possible because binomial limits must be integers. However, if you use sigma limits, the sigma limits produced are allowed to take on non-integer values.

p Chart Options

Here are a few specific comments about options for *p* charts:

Sigma. This is not an option for this chart.

Indicate subgroup size (AGG=TOTAL). With *p* and *u* charts, you can indicate that aggregate data are the total counts rather than the proportions.

Control limits. The default control limits are probability limits using the binomial distribution of sample proportions for a given sample size. If you specify sigma limits, the numbers are taken to be scaled in units of the standard error of binomial proportions for each sample.

Input data. The *y* variable value must be either 0 or 1. Aggregate by subgroup signals that input data are already aggregated as proportions—one for each sample or subgroup. The Indicate subgroup size option signals that the input data are total counts rather than proportions.

c Charts

The *c* chart plots sample counts as a function of a sample identifier, using the Poisson distribution as a model. The center line for a *c* chart is the mean number of successes (or failures) per sample unit times the number of sample units involved for each sample. Poisson distributions are additive, so the expected value when a sample involves more (or less) than one sample unit is the expected count per unit times the number of sample units. See the discussion by Montgomery (2001) for more information. The expected rate per unit is estimated from the data as the mean count per sample unit. Control limits are Poisson probability limits for a given alpha once the value of the Poisson center line is known because the center line determines which Poisson distribution is being used. Thus, the control limits vary directly as a function of the center line value.

The Poisson distribution is discrete. Therefore, control limits are set to integer values, which means that the reported value for alpha may differ.

Sample size versus sample units. For most applications of the Poisson distribution to control charts, one generally counts the instances of an attribute within some temporal or spatial entity. For example, the number of assembly-line accidents occurring each month might be plotted as a function of months; or, the number of defects in equal-sized rolls of cloth might be counted and plotted as a function of roll number. The concept of sample size for the Poisson refers to the number of potential successes or failures within each sample. Theoretically, this is an infinitely large number. The assembly-line and cloth examples provide an extremely large (and equal) number of opportunities for success or failure within each sample.

For both Poisson charts (*c* and *u* charts), numbers given in a *weight* variable are assumed to be the number of sample units involved in the count of successes or failures, as opposed to the usual sample sizes. A sample unit is the spatial or temporal entity within which successes or failures are counted and within which the potential number of successes or failures could be extremely large (theoretically infinite). Each roll of cloth mentioned above constitutes one sample unit. If each sample in the file consists of one such unit, and each contains the same number of opportunities for success or failure, then the weight (that is, the number of sample units) is 1 for each sample. In this case, the WEIGHT option is not used because its default is 1.

Now consider a situation in which sample units do not each have the same number of opportunities for success or failure. For example, the rolls of cloth may each contain a different number of square yards of fabric. If sample units do not each provide the same potential number of successes or failures, your file should contain a variable that designates the number of sample units involved in each sample. You would use this variable in a WEIGHT command (Data menu) before running SHEWHART. First decide on a unit of measure. For example, the size of each roll of cloth might be designated in square yards, where each sample unit contained 1 square yard. A 500 square-yard roll would then have a weight value of 500. The unit of measure is arbitrary, however, so you could also designate the unit of measure as, say, 50 square yards of fabric. Then, a roll containing 500 square yards would have a weight value of 10. The weight variable can meaningfully take fractional values for sample units. For example, a 525 square-yard roll would have a weight of 10.5 and a 40 square-yard roll would have a weight value of 0.8 (if one sample unit were defined as 50 square yards).

Note that the only case where non-integer weight values are meaningful is when you are producing Poisson *c* and *u* charts with data aggregated by subgroup. In all other instances, values are truncated to integers and used to indicate replications of cases.

c Chart Options

Here are a few specific comments about options for *c* charts:

Sigma. This is not an option for this chart.

Control limits. The default control limits are the default probability limits using the Poisson distribution with a given mean size. If you specify preset limits, they are placed at whatever value(s) you specify. You can use sigma limits to display traditional sigma limits for the chart.

Input data. Raw data must be 0's or 1's. Aggregate by subgroup (AGG) indicates that input data are already aggregated as counts, one for each sample or subgroup. The sample size of the Poisson distribution is, in theory, infinitely large. Thus, FREQUENCY is not used to indicate sample size in the usual sense. Frequency reads the data as though any given case were present *n* times, where *n* is the value of the weight variable. If *n* is not an integer, it is truncated. This is also true for raw input data for the *c* chart. Do not use FREQUENCY with Aggregate by subgroup unless you intend that each aggregate (count) is to be replicated as many times as the frequency variable indicates.

u Charts

The *u* chart plots a count per sample unit as a function of a sample identifier, using the Poisson distribution as a model. Everything in the discussion about *c* charts also applies for *u* charts. The major difference is that the *c* chart plots total counts regardless of the number of sample units per sample, while the *u* chart plots rate per sample unit. Thus, if each sample contains only one sample unit, the *c* and *u* charts are the same.

The *c* and *u* charts may produce empirical values of alpha that do not exactly match the value specified, just as the binomial charts. This is due to the discrete nature of the Poisson distribution, which requires that limits be integers.

A word of caution regarding the distinctions between data types and the Indicate subgroup size (AGG=TOTAL) versus Aggregate by subgroup (AGG) for both the binomial and Poisson charts: These options give you a lot of flexibility in analyzing data but can also lead to confusion. For example, you could mistakenly use one of these options with raw data and receive no error message. Mixing and matching options with data types and shifting between *c* and *u* charts can get confusing and lead to strange charts.

u Chart Options

Here are a few specific comments about the options for *u* charts:

Sigma. This is not an option for this chart.

Control limits. The default control limits are probability limits using the Poisson distribution with a given mean.

Aggregate by subgroup (AGG). Signals that input data in the *y* variable are already aggregated as rate per unit, one rate for each sample or subgroup. The sample size for the Poisson distribution is, in theory, infinitely large.

Indicate subgroup size (AGG=TOTAL). Signals that the input data are total counts for each unit (sample or subgroup) rather than rates.

OC and ARL curves

SYSTAT can produce a graph of either an operating characteristic (OC) curve or the corresponding average-run-length (ARL) curve for each of eight statistical distributions. The OC curve shows the value of beta—the probability of a Type II error—as a function of a range of possible expected values of the appropriate sampling distribution. The ARL curve shows 1/(1–beta) as a function of a range of possible expected values of the sampling distribution. The ARL curve indicates the average number of runs (trials) one would expect to occur before an out-of-control signal was encountered, given various possible "true" values of a parameter. Both the OC and ARL curves depend on a specific center (null hypothesis) value and a specific set of control limits.

The standard deviation of a hypothetical population of individual observations is either implicitly or explicitly involved in the sampling distributions whose OC or ARL functions are being explored. Sigma refers to the standard deviation of the population of individual observations rather than the standard deviation of the sampling distribution for the statistic whose expected values are being plotted.

For example, the standard deviation of the sampling distribution of sample means—usually called the standard error—is the population standard deviation divided by the square root of the sample size. The Sigma option lets you specify the numerator of the standard error; you specify the population standard deviation rather than the standard error itself. This is true for each of the continuous distributions that are denoted by the *X*, variance, *s*, and *R* types. For each, the standard deviation (standard error) of the

sampling distribution is computed from the population standard deviation (sigma) and the sample size (*N*).

For all distribution types and for all possible values of the parameter being plotted, beta is always defined the same way:

- For a two-tailed alpha, beta is the proportion of the distribution that is equal to or greater than the lower limit *and* less than or equal to the upper limit.
- For an upper-tail alpha, beta is the proportion of the distribution that is less than or equal to the upper control limit.
- For a lower-tail alpha, beta is the proportion of the distribution that is greater than or equal to the lower control limit.

Setting probability limits (PLIMITS) to 0.05 will place a limit (critical value) in the *lower* tail of a distribution; setting probability limits to 0.95 will place one in the *upper* tail of the distribution; probability limits of 0.05, and 0.95 will place critical values (limits) in *both* tails.

The computation for continuous distributions would not change if we said "less than" or "greater than" rather than "less than or equal to" and "greater than or equal to" in the previous sentences. However, this distinction can seriously affect the computation for discrete distributions with small sample sizes.

Operating Characteristic Curves

SYSTAT plots an operating characteristic (OC) curve for any of eight statistical distributions, showing the probability of a Type II error (beta) as a function of a range of possible expected values of the sampling distribution. The distribution used depends on the chart type.

Operating Characteristic Curve Dialog Box

To open the Operating Characteristic Curve dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      Operating Characteristic Curve...

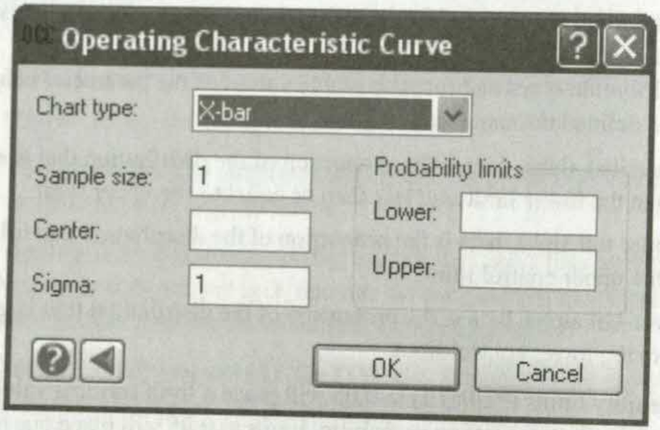


Chart type. Available types include *X*-bar, variance, *s*, *R*, *np*, *p*, *c*, and *u*. The *X*-bar type produces an OC curve for a normal distribution plotted as a function of a range of possible values for the population mean. For variance, you get a curve for a range of possible population variances, using the chi-square distribution as a model for the curve. For *s*, you get the expected standard deviation, again using the chi-square distribution as a model. *R* produces the OC curve for the expected range of a normally distributed variable. *np* produces the OC curve for the binomial count, and *p* produces curves for the binomial proportion. *c* produces charts for the Poisson count, and *u* produces curves for the Poisson rate per sample unit.

Sample size. The sample size.

Center. An *a priori* value to use for the center line of the chart.

Sigma. An *a priori* value for the population within-sample standard deviation.

Probability limits. The proportion of the sample statistic values expected to be below each control limit.

Average Run Length Curves

SYSTAT plots an average-run-length (ARL) curve for any of eight statistical distributions, showing 1/(1–beta) as a function of a range of possible expected values of the sampling distribution. The distribution used depends on the chart type.

Average Run Length Dialog Box

To open the Average Run Length dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      Average Run Length...

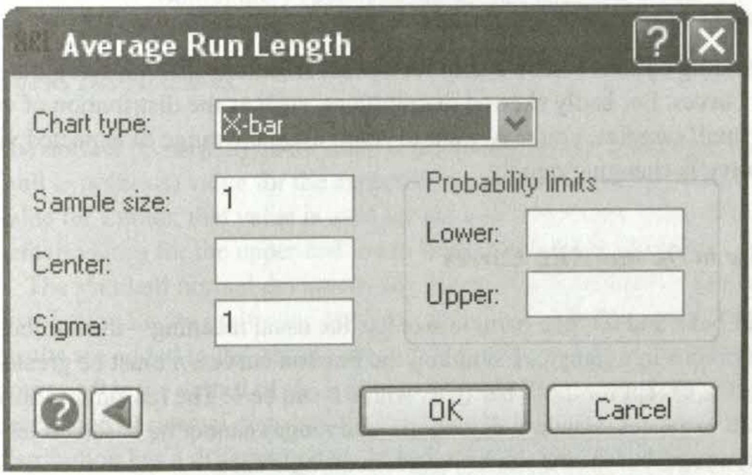


Chart type. Available types include *X*-bar, variance, *s, R, np, p, c* and *u*. The *X*-bar type produces an ARL curve for a normal distribution plotted as a function of a range of possible values for the population mean. For variance, you get a curve for a range of possible population variances, using the chi-square distribution as a model for the curve. For *s*, you get the expected standard deviation, again using the chi-square distribution as a model. *R* produces the ARL curve for the expected range of a normally distributed variable. *np* produces the ARL curve for the binomial count, and *p* produces curves for the binomial proportion. *c* produces charts for the Poisson count, and *u* produces curves for the Poisson rate per sample unit.

Sample size. The sample size.

Center. An *a priori* value to use for the center line of the chart.

Sigma. An *a priori* value for the population within-sample standard deviation.

Probability limits. The proportion of the sample statistic values expected to be below each control limit.

Scaling in OC and ARL Curves

By default, SYSTAT sets a scale for both axes of an OC or ARL plot. It also uses a rounding facility to make the tick marks on the axes round numbers. If you specify both a minimum and maximum for an axis (available with commands), then this rounding facility is bypassed. If you specify only a minimum, the minimum of the axis is set to this value, and the rounding facility is used to scale the axis. If you specify only a maximum, the program comes as close as it can to honoring this maximum value while scaling and rounding to get the axis values.

The scaling options can be useful for certain situations that arise in plots of OC and ARL curves. For badly skewed distributions, such as the distribution of variance for very small samples, you may want to zoom in on the range of expected values where the curve is changing rapidly.

Sample Size in OC and ARL Curves

For all but *c* and *u* types, sample size has the usual meaning—the number of observations in a sample. Excluding the Poisson curves, *n* must be greater than 1 for all curves except for the *X*-bar type, where it can be 1. The reason for this is obvious—sample variances, standard deviations, and ranges cannot be based on less than two observations, but means can be based on a single case.

For the Poisson (*c* and *u* type) curves, *n* is the number of sample units involved. A sample unit is the temporal or spatial entity in which instances of the attribute being monitored can occur; for example, a month, a square meter, etc. Theoretically, the sample size is infinitely large within a sample unit. In practice, though, while the number of opportunities for an event to occur is large within a sample unit, it is not infinitely large, and sample units could differ in size. Large rolls of cloth, for example, could differ in the number of square meters per roll, but the number of opportunities for defects would still be large in any selected square meter. The sample unit, in this case, could be a square meter, with different rolls having different numbers of square meters, or sample units. Thus, for *c* or *u* type curves, we specify the number of sample units over which a count of an attribute is to be made for *n*. Note that *n* here could have a non-integer value.

All of this said, it does not really make any difference how many sample units are involved for a *c* type curve, which involves just the total count of events occurring over all sample units. The total count is the total count, regardless of how many sample units are involved, so you get the same curve regardless of the value of *n*. That is, we are just

dealing with an OC or ARL curve for a simple Poisson distribution, the mean of which represents the expected count over all sample units.

The *u* type curve, however, is a different matter. Because this curve is based on rate per sample unit, the number of sample units involved affects the curve. Here we are dealing with a Poisson distribution, the mean of which is the expected rate per sample unit. All else constant, the more the sample units that are involved, the more like a normal distribution curve the result will be. If $n = 1$, then the *c* and *u* curves will be identical. See Montgomery (2001) for more details on this topic.

Continuous Distributions

The normal (Gaussian) distribution is the model for the *X*-bar type curves. The center (null hypothesis) value for the expected value (mean) is 0 by default. If you specify a value for Center, that value is used for the null hypothesis value of the mean. The default values for the upper and lower limits are found as follows:

The standard normal deviations (Z) below which probability limits p_1, p_2 of the distribution lie are multiplied by sigma and divided by the square root of *n*. These results are added to the center value; the resulting sums are the values of the control limits. After the values of the limits are known, beta is computed for each of a succession of normal distributions, each having the same standard error as the null distribution but a different mean. Then beta is plotted as a function of these possible means.

The chi-square distribution with *n* degrees of freedom is the model for variance type curves. If you specify Center, the null hypothesis value used for the population variance is the square of that value. If Center is not specified, the squared value of Sigma is used. If neither Center nor Sigma is specified, then 1 is used as the null hypothesis value for Var (You cannot specify both Center and Sigma for variance curves). The default limits are the values below which probability limits p_1, p_2 of the chi-square distribution lie are multiplied by the null hypothesis value of Var and divided by *n*-1, where *n* is the sample size. After the values of the limits are known, beta is computed for each of a succession of unstandardized chi-square distributions. Each is the standardized chi-square distribution, with *n*-1 (degrees of freedom) multiplied by a different possible value of Var (the population variance) and divided by *n*-1. Then beta is plotted as a function of these population variances.

The *s* type curves are computed similar to the variance curves, also using the chi-square distribution as a model. If Center is specified, the null hypothesis value of the population variance is computed as the square of center/c_4, where c_4 is a correction for bias based on sample size (The expected value of the distribution of sample standard

deviations is c_4 times the population standard deviation). See Montgomery (2001) for the mathematical definition of c_4. If Sigma is specified, then its square is used as the null hypothesis value of the population variance. If neither Sigma nor Center is specified, then 1 is used as the population variance. Control limits for the unstandardized chi-square distribution are found in the same way as for the variance curves discussed above, and beta is computed in the same way as for those curves. But beta is plotted as a function of each of a set of possible expected values of the sampling distribution of *s*, not variance. These expected values are c_4 times each of a set of possible population standard deviations.

The *R* type curves are based on the sampling distribution of the range for a standard normal distribution, given the sample size (*n*). Let us call this standardized range *W* to distinguish it from *R*, the unstandardized range (*W* is also called the relative range). The expected value of the *W* distribution multiplied by sigma is the expected range for a normal distribution with a standard deviation of sigma. Thus, SYSTAT must be directly or indirectly given a value for sigma. An indirect way to supply it is to specify a value for Center; then sigma is computed as center/E(*W*), where E(*W*) is the expected standardized range for the given *n*. (SYSTAT automatically computes E(*W*) in this case; you do not have to supply it.) You can directly specify a sigma with the Sigma option. You cannot specify both Center and Sigma. If neither is specified, sigma is set to 1. Control limits are found for the null hypothesis value of the range, which is either the specified center or the center computed as sigma times E(*W*). After these limits are found, beta is computed for each of a set of possible values of the expected range, E(*W*) times sigma. Then beta is plotted as a function of these expected values.

Discrete Distributions

The *np* and *p* chart types are based on the binomial distribution, and the *c* and *u* types are based on the Poisson distribution. As stated earlier, the definition of control limits for discrete distributions must be carefully stated and observed for precise OC and ARL results.

Control limits are usually defined so that a data point that falls right at the limit is considered "in control." Only points that strictly exceed the limit are "out of control." This concern about points right at the limit makes no theoretical difference when one is dealing with a truly continuous statistic. However, the precise definition is important for discrete distributions. The true values of alpha or beta can be severely affected if one treats a sample value that is right at the limit as being out of control rather than in control.

For *np* and *c* type curves, control limits are always placed at integer values because outcomes for the two distributions must be integers. The rules used to compute control limits for these OC and ARL curves follow:

Let us define alpha as $1\text{-}p$ if probability limits = p and $p > 0.5$; or as just p if probability limits = p and p is less than or equal to 0.5; or as $p_1+1\text{-}p2$ if probability limits = p_1, p_2.

For an upper-tail alpha value, SYSTAT sets the upper control limit to the integer beyond which no more than alpha of the distribution lies. For a lower-tail alpha value, SYSTAT sets the lower control limit to the integer below which no more than alpha of the distribution lies. When both upper and lower control limits are used, the same rules apply except that alpha is replaced with alpha/2 in the previous sentences. Notice that for small samples or for badly skewed distributions, the lower limit may end up at 0, and/or the upper limit may be the sample size for a binomial distribution. If both of these conditions exist for a two-tailed alpha, no sample can ever be out of control because no count of a sample attribute can be less than 0 or more than the sample size. In this case, all values of beta will be 1.

For example, this will happen if you use the default, two-tailed alpha (0.0027) for an *np* curve when the sample size is 2 and the population *p value* is 0.5. For this binomial distribution, the probability of a count of exactly 2 is 0.25. Thus, if one treats a count of 2 as out of control, the true upper-tail alpha would be 0.25. This is greater than the desired alpha, so SYSTAT sets the upper limit at exactly 2, meaning that counts that are greater than 2 are considered out of control. The sample size is 2, so no sample count can exceed 2. A similar argument applies to the lower limit. The probability of a count of exactly 0 for the distribution in question is also 0.25. To avoid exceeding the alpha value, the lower limit is set at 0, meaning that counts of 0 are considered in control. The upper and lower limits found here demand that no sample be rejected as out of control. The true value of alpha used by SYSTAT in this case would be 0, and beta would be 1 for any possible value of *np* because no count can possibly fall outside the control limits.

Similar procedures apply to the distributions for the *p* and *u* curves. These distributions must also be discrete because they are forms of the binomial and Poisson, respectively. The only difference for them is that the integer value found for the proper binomial or Poisson is converted to a proportion or rate. Thus, only a finite set of limits is possible. There are, of course, other approaches to setting control limits for the discrete distributions. The appeal of the approach taken here is that the actual probability of a Type I error, alpha, never exceeds the specified value, although it might be less than that value.

The exact value SYSTAT uses for alpha in the case of a discrete distribution can be found by locating the beta value corresponding to the null hypothesis point on the OC curve. The actual alpha is 1 minus this value.

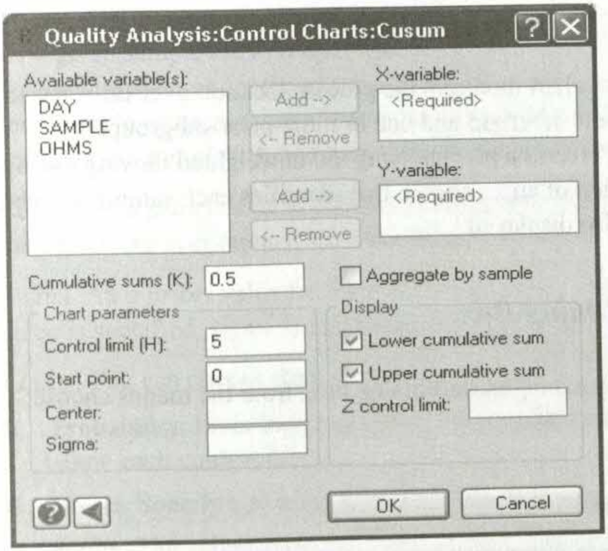
Cusum Charts

Cusum produces upper and lower cumulative sum charts for a *y* variable as a function of the *x* variable that identifies each sample. Cumulative sum charts can be produced by using various methods, some easier to understand than others. The methods used here are those suggested by Ryan (2000), who says, "If quality control personnel are to make the transition from *X*-bar charts to Cusum procedures, it seems reasonable to assume that they will be drawn toward procedures that are easy to understand and that bear some relationship to an *X*-bar chart." Ryan presents the methods of Lucas (1982) and Lucas and Crozier (1982); the same methods are used here.

Cumulative Sum Chart Dialog Box

To open the Cumulative Sum Chart dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      Cusum...



X-variable. Select the variable you want to choose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Cumulative sums (K). Uses the specified value to compute cumulative sums.

Control limit (H). Control limit for the cumulative sum.

Start point. Start for initial sum.

Center. An *a priori* center value. The default value of center is the grand mean of all of the data.

Sigma. An *a priori* value for the population standard deviation.

Aggregate by sample. Indicates that input data are already aggregated by sample. If you select this option you should use FREQUENCY (Data menu) to indicate the sample size. You must also supply a value for sigma because SYSTAT cannot compute sigma from aggregated data.

Lower and Upper cumulative sum. You can plot the upper chart, lower chart, or both.

Z control limit. If output length is set to Long, this option flags cases whose absolute individual *Z* values exceed the specified value in the tabular output. (The value is converted to an absolute value internally.) This option has no effect on the cumulative sum plot itself.

Moving Average Charts

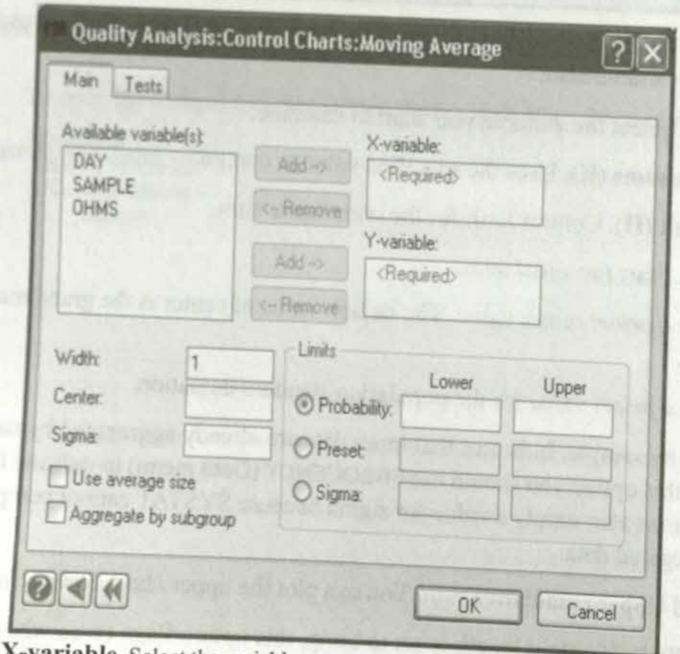
The moving average chart monitors the process location over time, based on the average of the current subgroup and one or more prior subgroups.

Moving average creates a plot showing the unweighted moving average of a *y* variable, as a function of an *x* variable that identifies each sample. A center line and control limits are also displayed.

Moving Average Chart Dialog Box

To open the Moving Average Chart dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      Moving Average...



X-variable. Select the variable you want to choose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Width. The number of samples over which the moving average is computed. The data for a given sample are averaged with the data from the previous *n-1* samples to obtain the moving average for the given sample. The number of cases used to compute limits for a sample is the number of cases in the sample plus the number of cases in the previous *n-1* samples. A width of 1 corresponds to a Shewhart *X*-bar chart.

Center. An *a priori* value to use for the center line of the chart. If no value is specified, the center line is computed from the data.

Sigma. An *a priori* value for the population within-sample standard deviation. If no value is specified, sigma is computed from the data.

Limits. You can choose from Probability, Preset, or Sigma limits.

- **Probability.** Enter the proportion of the sample statistic values expected to be below each control limit.
- **Preset.** Specify a *priori* values for control limits.
- **Sigma.** The values specified are assumed to be in units of the standard error of the statistic being plotted.

Use average size. Control limits are computed using average subgroup size for each subgroup. Use this option with caution. You will get a warning message if the size of any sample deviates by more than 10% from the average sample size.

Aggregate by subgroup. Indicates that input data are already aggregated by subgroup. In this case you should use FREQUENCY (Data menu) to indicate subgroup sizes.

After constructing the chart, you can make suitable interpretations of it by clicking Tests in the Moving Average Chart dialog box. Eight run tests are available here. These tests indicate if the process is in control or out of control depending on the tests you have selected (For details about the tests refer to the Tests section).

By default, the center, sigma, and probability limits are computed from the raw data. The default center line is the simple average of all of the values of the *y* variable. This value is shown on the output as *Estimated Population Mean*. The default value for sigma is the square root of the weighted mean of all of the within-sample variances, where each weight is an individual sample size minus one. This value is shown on the output as *Estimated Population Standard Deviation*. The default values of limits are computed as the standard normal deviates that exclude alpha of the standard normal distribution, multiplied by sigma and divided by the square root of the number of cases on which the moving average was based. Therefore, the "sample size" on which the control limits are based is a function of the value of Width.

Some information about the default computation for sigma bears repeating here. First, you must use raw data for this computation to occur (that is, Aggregate by subgroup cannot be used). The program first computes the sums of squared deviations around the sample mean for each sample independently. These sums of squares are summed over all samples and then divided by *n-k*, where *n* is the total number of data points in all samples being used, and *k* is the number of samples being used. The resulting number is an unbiased estimate of the population variance. The square root of this number is used as the default estimate of the population standard deviation.

Exponentially Weighted Moving Average Charts

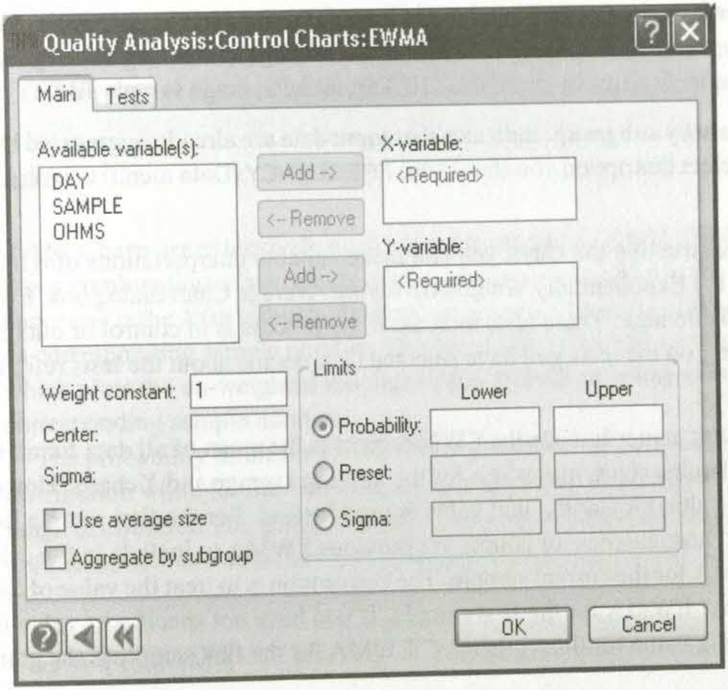
A control chart based on an exponentially weighted moving average (EWMA) was presented by Roberts (1959) and is also described by Montgomery (2001). The EWMA chart plots the exponentially weighted moving average of the *y* variable as a function of an x variable that identifies each sample. The plot has a center line and control limits.

The EWMA is also called the geometric moving average (GMA) chart. A weighting constant, *K*, is used, such that each successive sample mean is multiplied by *K* and added to the previous value of the EWMA times *1-K*. The resulting new value of the EWMA is then plotted as a function of the sample identifier.

Exponentially Weighted Moving Average Chart Dialog Box

To open the Exponentially Weighted Moving Average Chart dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      EWMA...



X -variable. Select the variable you want to choose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Weight constant. The weight constant for computing the exponentially weighted moving average. The value must be between 0 and 1.

Center. An *a priori* value to use for the center line of the chart. If no value is specified, the center line is computed from the data.

Sigma. An *a priori* value for the population within-sample standard deviation. If no value is specified, sigma is computed from the data.

Limits. You can choose from Probability, Preset, or Sigma limits.

- **Probability.** Enter the proportion of the sample statistic values expected to be below each control limit.
- **Preset.** Specify *a priori* values for control limits.
- **Sigma.** The values specified are assumed to be in units of the standard error of the statistic being plotted.

Use average size. Control limits are computed using average subgroup size for each subgroup. Use this option with caution. You will get a warning message if the size of any sample deviates by more than 10% from the average sample size.

Aggregate by subgroup. Indicates that input data are already aggregated by subgroup. If you select this option you should use FREQUENCY (Data menu) to indicate subgroup sizes.

After constructing the chart, you can make suitable interpretations of it by clicking Tests in the Exponentially Weighted Moving Average Chart dialog box. Eight run tests are available here. These tests indicate if the process is in control or out of control depending on the tests you have selected (For details about the tests refer to the Tests section).

The default center line for the EWMA chart is the mean of all data for all samples included in the chart, just as it is for the moving average and X charts. However, if you assign a value for Center, that value is used instead. For the first sample listed in an EWMA chart, there is, of course, no previous EWMA to include in the calculation of the EWMA for the current sample. The convention is to treat the value of Center as the "previous" EWMA for the first sample. If you have not specified a value for Center, the starting value for this "previous" EWMA for the first sample is the grand mean of all data.

As always, the value of sigma is the population standard deviation of the individual data values within samples. If you do not assign a value to Sigma, it is computed from the data (See the discussion on moving average charts for details.). The standard deviation (standard error) of the sampling distribution of the EWMA is given by:

$$\text{S(EWMA)} = \frac{\text{SIGMA}}{\sqrt{n_t}} * \sqrt{\frac{K}{2-K} * (1-(1-K)^{2t})}$$

where S(EWMA) is the standard error that we seek, n is the sample size, K is the weighting constant, and t is the sample number. Note that the sample size could be different from sample to sample. This formula results in an S(EWMA) that increases for each successive sample, but the rate of increase decreases for successive samples. The control limits are based on S(EWMA), so they are narrower for the earlier samples than for later samples, as seen in the example.

The default upper control limit (UCL) is computed by multiplying the value of the standard normal deviate that has $1-p_2$ of the normal distribution above it by S(EWMA) and added to Center to get the default UCL. Similarly, the lower control limit is the

standard normal deviate corresponding to p_1 times S(EWMA) added to the value of Center. Here, p_1 and p_2 are values specified for probability limits. As for all of the SYSTAT control charts, if you specify preset limits, your limit values are used directly.

X-MR Charts

X-MR Charts are extensively used when the sample available consists of a single unit. It is a combination of two charts drawn one above the other in the same frame. The upper one is the *X* (or individual) chart, which plots the value of the sample unit against its corresponding sample number. The lower one is the MR (or moving range) chart, which plots the un-weighted moving ranges of each sampling unit against the corresponding sample number.

The probability limits for the *X* chart in SYSTAT are computed assuming a normal distribution while for the MR chart the population is assumed to follow a Studentized range distribution.The general form for the control limits for the *X* and the MR chart are shown below: For the probability limits with values P_L and P_U, sigma is estimated by:

$$\hat{\sigma} = (1/d_2)\overline{MR}$$

if the estimation is done through the mean method or,

$$\hat{\sigma} = (1/d_4)MR_{(med))}$$

if the estimation is done by the median method.

The upper control limits of the MR chart in SYSTAT are calculated by multiplying the estimated value of sigma by the Studentized inverse of P_U. Similarly the lower limit is calculated by multiplying the estimated sigma by the Studentized inverse of P_L. The control limits for the *X* chart are given by:

$$UCL = \text{estimated mean (x)} + \hat{S}\Phi^{-1}(P_U)$$

$$LCL = \text{estimated mean (x)} - \hat{S}\Phi^{-1}(P_L)$$

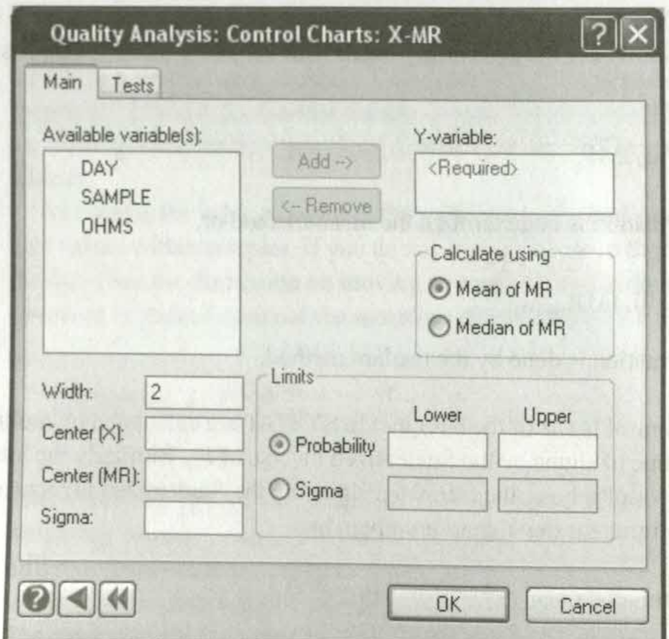
where $\hat{S}$ is the estimated value of the population standard deviation (*S*) of the X values and $F^{-1}(.)$ is the inverse of the standard normal distribution function. The value of $\hat{S}$ depends on the estimation method applied. The values of *d*'s are obtained from tables

corresponding to the value of the width specified. When the sigma limits are specified, the control limits are also calculated in the similar manner but the computation formula is different. If no limits are chosen as such, the default limits are, as usual, the probability limits. For both the charts, SYSTAT displays the center lines and the corresponding control limits. This chart is also known as an I-MR chart.

X-MR Chart Dialog Box

To open the X-MR Chart dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      X-MR...



Y-variable. Select the variable you want to examine.

Width. The number of samples over which the moving range is computed. The range of the previous (*k-1*) values along with the current value is calculated to obtain the

moving range of width *k* for the given sample. The number of cases used to compute limits for a sample is the number of cases in the sample plus the number of cases in the previous (*n-1*) samples. A width of 2 is the default.

Center (X). An *a priori* value to use for the center line of the *X*-chart. If no value is specified, the center line is computed from the data, which is the grand mean of the *X*-variable values.

Center (MR). An *a priori* value to use for the center line of the MR-chart. If no value is specified, the center line is computed from the data, which is the mean or the median of the moving ranges. You can give any non-negative numeric value here.

Sigma. An *a priori* value for the population within-sample standard deviation *s*. If no value is specified, *s* is computed from the data. The default value for *s* is the standard deviation of the *X*-variables. This value is shown in the output as *Estimated Population Standard Deviation.*

Limits. You can choose from Probability or Sigma limits. The default values of limits are computed as the standard normal deviates that exclude alpha of the standard normal distribution, multiplied by sigma and divided by the square root of the number of cases on which the moving average was based. Therefore, the "sample size" on which the control limits are based is a function of the value of Width.

- **Probability.** Enter the proportion of the sample statistic values to be below each control limit.
- **Sigma.** The values specified are assumed to be in units of the standard error of the statistic being plotted.

Calculate using. Two calculation options are available here. They are:

- **Mean of MR.** By default the mean of the moving ranges is used as the statistic to calculate the control limits for the X and the MR charts.
- **Median of MR.** The median of the moving ranges is used as the statistic to calculate the control limits for the X and the MR charts.

After constructing the chart, you can make suitable interpretations of it by clicking Tests in the X-MR Chart dialog box. Eight run tests are available here. These tests indicate if the process is in control or out of control depending on the tests you have selected. (For details about the tests refer to the Tests section.)

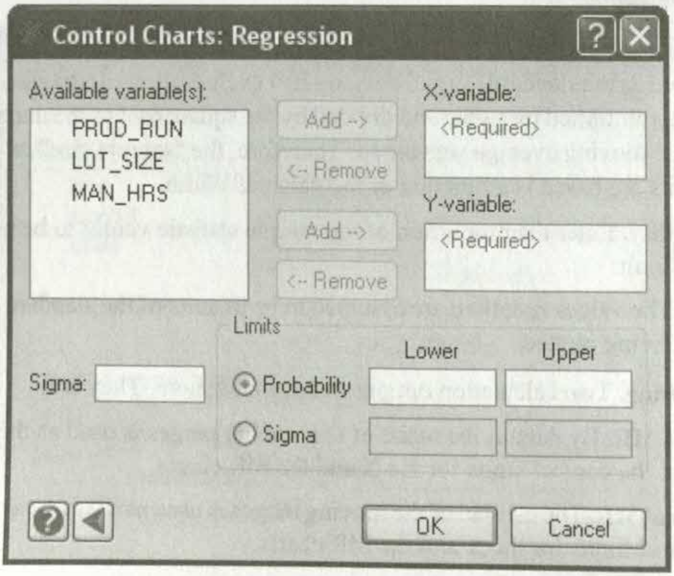
Regression Charts

A regression chart shows a *y* variable, a regression line, and prediction limits as a function of an *x* variable.

Regression Chart Dialog Box

To open the Regression Chart dialog box, from the menus choose:

Advanced
 Quality Analysis
  Control Charts
   Regression...



X-variable. Select the variable you want to choose as *x* variable.

Y-variable. Select the variable you want to examine.

Sigma. The standard error of the estimate for the regression line. If not specified, the regression line and sigma are calculated from the data.

Limits. You can choose from Probability or Sigma limits.

- **Probability.** Enter the proportion of the sample statistic values expected to be below each control limit.
- **Sigma.** The values specified are assumed to be in units of the standard error of the statistic being plotted.

The default control limits for the regression chart are prediction limits for individual instances of data. They are not the confidence limits that you are used to seeing for the predicted value of the *y* variable. See Neter et al. (1996) for a discussion of this distinction. The equations for these prediction limits are:

$$\text{UCL} = \text{Predicted } y + t * \text{Sigma} * \text{Q}$$

and

$$\text{LCL} = \text{Predicted } y - t * \text{Sigma} * \text{Q}$$

where t is the t-distribution deviate that corresponds to alpha or alpha/2, and

$$Q = \sqrt{\left(1 + \frac{1}{N}\right) + \frac{(Y_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}}$$

One further note about the Sigma option: First, notice that Sigma is used as the population standard error of estimate, not as the population standard deviation of the *y* variable. If Sigma is specified, then the population value of the standard error of estimate is assumed to be that value. When the population value of sigma is known, the *t* distribution is not appropriate. In this event, the program computes the default prediction limits from normal distribution deviates instead of *t*-distribution deviates.

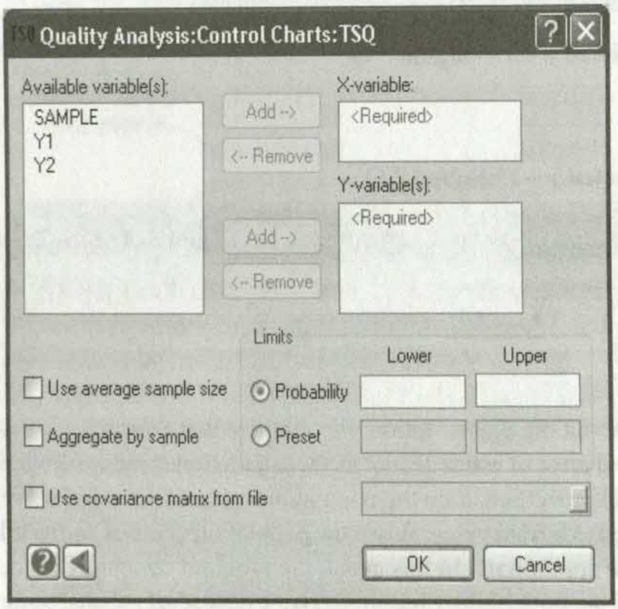
TSQ Charts

TSQ creates a control chart showing Hotelling's T^2 for up to 10 *y* variables as a function of an *x* variable that identifies each sample. Use these charts to analyze data where multiple, possibly correlated, variables must be controlled. See Ryan (2000) or Montgomery (2001) for details.

TSQ Chart Dialog Box

To open the TSQ Chart dialog box, from the menus choose:

Advanced
  Quality Analysis
    Control Charts
      TSQ...



X-variable. Select the variable you want to choose as *x* variable. You can use a categorical variable also.

Y-variable. Select the variable you want to examine.

Use average sample size. Uses the average sample size as the sample size for each sample.

Aggregate by sample. Indicates that input data for each sample consist of sample means for each of the *y* variables. If you select this option you must use FREQUENCY (Data menu) to specify the sample sizes. You must also supply a covariance matrix because the program cannot compute the variance-covariance matrix from aggregated data.

Use covariance matrix from file. Specify the name of a file containing an *a priori* variance-covariance matrix and vector of means. If no file is specified, SYSTAT computes the pooled variance-covariance matrix (not if it is a case of aggregated data) and the grand means from the data.

Limits. You can specify probability or preset limits.

- **Probability.** Enter the proportion of the sample statistic values expected to be below each control limit.
- **Preset.** Specify *a priori* values for control limits.

The computation of the empirical *T*-squared for each sample is straightforward and conforms to the usual formula given in most textbooks. However, the computation of the UCL requires some discussion because there is some variability among authors as to the recommended procedure. The formula used in SYSTAT is algebraically equivalent to one given by Ryan (2000):

$$\text{UCL} = \frac{p(n-1)(k-1)}{(kn-k-p+1)} * F_{\alpha,(p,kn-k-p+1)}$$

where p is the number of dependent variables, n is sample size, k is the number of samples, and F is the F distribution value that has alpha of the distribution above it, with p and kn-k-p+1 degrees of freedom. This formula is appropriate when the variances, covariances, and grand means have been calculated from the data. Other formulas, also given by Ryan, are appropriate for situations where a set of data has been used to establish an upper control limit for future sample values of *T*-squared, or where the population means are known. There is no way for SYSTAT to know which of these assumptions, or combinations of them, apply in a given context. However, the LIMITS option allows you to establish a value for the upper control limit on whatever basis you want, using whatever formula you want. The basic data for alternative formulae are in the SYSTAT output, and you can use the CALCULATE command to get any F values you need.

Process Capability Analysis

A primary objective of a quality-conscious manufacturer is to produce a product a very large percentage of which will meet the specifications, thereby reducing defective products to a small percentage. However, actual product attributes can fall outside acceptable tolerance specifications in certain manufacturing environments, on some

aspects of which at least the manufacturer can exercise control. A statistical tool which helps in the understanding of the inherent capability of a manufacturing process vis. a vis. the current level of process performance is Process Capability Analysis. This analysis focuses on the variability of a process so that unacceptable tolerances and assignable causes of poor performance in the system can be identified and corrected, and the process improved to deliver products to its capability.

The capability of a process is an indication of how closely it performs relative to the specifications, taking into account the inherent variability due to unavoidable causes. The capability is measured by a variety of indices, which are discussed in detail in this section. Some of these indices are based on the assumption that the quality characteristic concerned, usually continuous in nature, has a normal distribution. When this is so, a high percentage of the process measurements fall between $\pm 3\sigma$ of the process mean or center. That is, approximately 0.27% of the measurements would naturally fall outside the $\pm 3\sigma$ limits and the remaining (approximately 99.73%) would be within the limits. Since the process limits extend from -3σ to $+3\sigma$, the total spread amounts to about 6σ total variation. To make it more general we will use $k\sigma$ tolerance as the amount of process spread. Process performance indices use the total variation of the process (as opposed to only inherent variation in Process Capability).

Process Capability Indices

For a specific process, USL is the upper specification limit and LSL is the lower specification limit. The target value for the process is *T*. You can estimate the mean (μ) and variance (σ^2) for the measured characteristics. Here *k* is used to compute process spread. On the basis of the assumption of normal distribution of the characteristic, the most commonly used performance indices (Pp, PPU, PPL, Ppk, Ppm, and Pmk) are expressed below (see Breyfogle, 2003 for more details).

$$P_p = \frac{USL - LSL}{k\sigma}$$

$$PPU = \frac{2(USL - \mu)}{k\sigma}$$

$$PPL = \frac{2(\mu - LSL)}{k\,\sigma}$$

$$P_{pk} = \min(PPU, PPL)$$

$$P_{pm} = \frac{USL\text{-}LSL}{k\sqrt{(\mu\text{-}T)^2 - \sigma^2}}$$

$$Pmk = 2\min\left(\frac{(T - LSL) - |T - \mu|}{k\sigma}, \frac{(USL - T) - |T - \mu|}{k\sigma}\right)$$

Box-Cox Power Transformation

The expressions for process capability indices given above are appropriate when the data follow a normal distribution. When the data arise from a non-normal distribution, we can use Box and Cox (1964) power transformation to transform the distribution of the data to normal distribution, and then we can use the above technique to compute these indices for the transformed data.

The Box-Cox transformation is as follows:

$$X(\lambda) = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(X), & \lambda = 0 \end{cases}$$

where X (>0) is the response variable. Using a maximum likelihood method, λ is estimated and data are transformed accordingly.

Process Performance vs. Process Capability

Depending on the sigma (σ) computation, which we use to get different process indices, we can refer to the indices as process performance or process capability indices. When the data set consists of multiple samples, we can compute two different indices of variability in the data, as follows. We can compute process indices using the overall variation in the process, ignoring the fact that data consists of multiple samples; the resulting indices are usually referred to as process performance indices. On the other hand, we can compute the indices by using some measure of inherent variability (say, within sample variability); in that case, the indices are called process capability indices.

Process Capability for Non-normal Data

Non-normality may occur in many practical situations. The capability measures mentioned above are not valid when the process characteristic does not follow a normal distribution. When data can be modeled by a known distribution form (e.g., Weibull distribution), we can follow the approach of replacing kσ by the length of the interval between the upper and lower 100*p* percentage points of the data. The *p* should be selected in such a way that the interval will cover a *k*σ interval for a normal (μ,σ) distribution. In particular for a 6σ-spread, we need to compute 0.135 and 99.865 percentiles from our assumed distribution. This methodology is applicable when all individual measurements are combined to determine a "long-term" capability. Both observed and expected "long-term" non-conformance rates can be determined using this methodology. See Cheng (1994-95), and Kotz and Lovelace (1998) for more details.

SYSTAT provides the following non-normal process performance indices:

$$P_p = \frac{USL - LSL}{Up - Lp}$$

$$PPL = \frac{Me - LSL}{Me - Lp}$$

$$PPU = \frac{USL - Me}{Up - Me}$$

$$P_{pk} = \min(PPL, PPU)$$

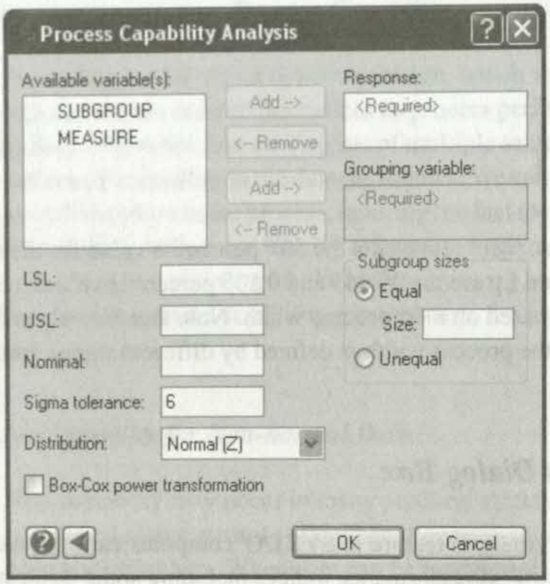
In the above expressions, 'Me' represents the 50th percentile value for the respective fitted distribution, Up and Lp are the 99.865 and 0.135 percentile values, respectively, if the computations are based on a 6σ process width. Note that the values for Up and Lp may be different, if the process width is defined by different *sigma* limits (e.g., ±2 times *sigma*).

Process Capability Analysis Dialog Box

The Process Capability Analysis feature in SYSTAT computes various process capability indices and process performance indices including some basic statistics. It also provides observed and expected process performance measure. You can perform the analysis for normal as well as for the following non-normal continuous distributions: beta, exponential, gamma, inverse Gaussian, lognormal, Rayleigh, and Weibull.

To open the Process Capability Analysis dialog box, from the menus choose:

Advanced
 Quality Analysis
  Process Capability Analysis...



Response. Select a variable you want to examine. It should be a continuous variable.

Grouping variable. A grouping variable contains a value that identifies group membership for each case. This is generally a categorical variable.

Subgroup sizes. If subgroup sizes are equal, then specify the Size. The sample size should be divisible by the selected subgroup size. For unequal subgroup sizes, select Unequal and add Grouping variable from the Available variable(s) list box.

LSL and USL. Specify the lower and upper specification limits. At least one is mandatory.

Nominal. You can specify a target value for the process. It is optional.

Sigma tolerance. Select a positive value for process spread. The default value is 6.

Distribution. Select a distribution from the drop-down list. The list includes the following distributions: Normal, Beta, Exponential, Gamma, Inverse Gaussian, Lognormal, Rayleigh, and Weibull. When your distribution choice is other than normal, only process performance indices are provided.

Box-Cox power transformation. You can use this option when data come from a non-normal distribution so as to make the distribution of transformed data near normal.

Using Commands

For Histogram, Pareto, Box-and-Whisker Plot, and Control Charts

- For a Shewhart chart

```
USE filename
QC
  SHEWHART yvar * xvar / TYPE = type TEST = 0 1 2 3 4 5 6 7 8
```

Replace *type* by XBAR, VAR, S, R, XBAR_S, XBAR_R, X, NP, P, C or U. You can either choose a specific TEST or perform all the tests. Default performs all the tests.

- For an OC or ARL curve:

```
QC
  OC / TYPE=type  N=n
```

Replace *type* by XBAR, VAR, S, R, X, NP, P, C or U. Replace *n* with the sample size (For an ARL curve, replace OC with ARL). Note that OC and ARL curves do not require data files.

- For other charts:

```
QC
  Command yvar * xvar / options
```

Replace *Command* by HIST, PARETO, BOX, RUNCHART, CUSUM, MA, EWMA, XMR, QCREGRESS or TSQ. For HIST specify an *xvar* while for XMR specify *yvar*.

The following options are available depending on the type of graph you create. Consult SYSTAT *Language Reference* for more information:

PLIMITS, LIMITS, SLIMITS, CENTER, SIGMA, P, CUM, BARS, BWIDTH, AVGN, AGG, AGG=TOTAL, YMIN, YMAX, XMIN, XMAX, K, H, WIDTH, MEAN, MEDIAN, MR, TREND, SHIFT, PATTERN, GROUP=GROUPVAR, ZCL,TEST.

- For Process Capability Analysis:

```
QC
PCA varname / USL=u LSL=l NOMINAL=m SIZE=size SIGMATOL=s DIST=
distribution BOXCOX
```

Replace *size* by *n* (or *varname* if the subgroup size is unequal). Replace distribution by NORMAL, BETA, EXPONENTIAL, GAMMA, INVERSEGAUSSIAN, LOGNORMAL, RAYLIEIGH or WEIBULL.

To SAVE a file, type:

```
SAVE filename
```

after QC command.

Usage Considerations

Types of data. QC uses only rectangular data. For OC and ARL curve no data file is required.

Print options. Control Charts provide Chart data as an extended output if the output length is set to MEDIUM or LONG. Process Capability Analysis offers two categories of output: SHORT (default) and MEDIUM or LONG. Besides process capability indices, MEDIUM and LONG output gives observed performance, expected capability and expected performance of the process.

Quick Graphs. In Pareto Charts, Box-and-Whisker Plot and Control charts, for each data set the corresponding graphs are plotted against the given *X*-variable. For Histogram, X-MR chart, Run Chart, only one variable is required to produce the corresponding Quick Graph. Process Capability Analysis produces a process capability chart with a histogram, probability density function and control lines.

Saving files. You can save the output of a control chart analysis into a file through commands. But the format of the saved file does not exactly match the output always. In particular for the output containing run tests the output will be slightly different.

BY groups. QC analyzes data by groups.

Case frequencies. In QC, you can specify a column for frequencies in Shewhart, Cusum, Moving Average, EWMA, Regression and TSQ charts to increase the number of cases.

Case weights. You can weight cases in QC by specifying a WEIGHT variable.

Examples

Example 1
Histogram

In this example, we use the *BOXES* data file, which contains raw data from Messina (1987). The ohms of electrical resistance in computer boxes were measured for five randomly selected boxes from each of 20 days of production. Thus, each *SAMPLE* contains five observations of resistance in ohms for each of 20 days.

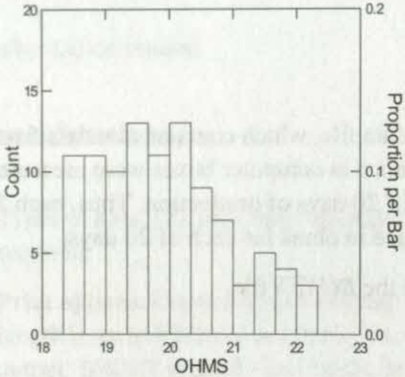
Here is a partial listing of the data in the *BOXES* file.

DAY	SAMPLE	OHMS
1.000	1.000	20.500
1.000	2.000	20.200
1.000	3.000	20.300
1.000	4.000	20.600
1.000	5.000	21.500
2.000	1.000	19.900
2.000	2.000	18.800
2.000	3.000	20.200
2.000	4.000	20.800
2.000	5.000	19.500
…	…	…
20.000	1.000	20.300
20.000	2.000	19.800
20.000	3.000	18.600
20.000	4.000	21.600
20.000	5.000	21.200

To obtain a histogram of these data, the input is:

```
QC
  USE BOXES
   HIST OHMS
```

The output is:



From the histogram, it is clear that the *OHMS* values in the Boxes data are distributed in the range 18 to 23. This distribution is neither uniform nor symmetric in nature. In fact it is slightly positively skewed (i.e., the frequencies of the smaller value of *OHMS* are more than the higher ones).

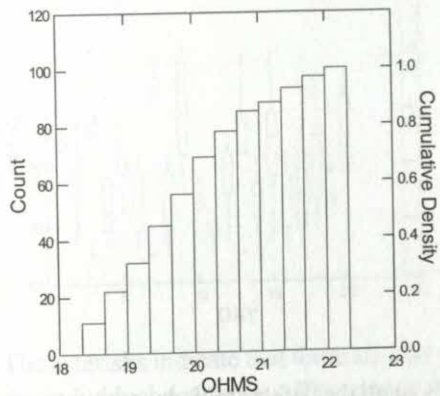
Example 2
Cumulative Histogram

To obtain a cumulative histogram of the same data set, the CUM option is used.

The input is:

```
QC
   USE BOXES
   HIST OHMS/ CUM
```

The output is:



From the cumulative histogram, we can infer that the marginal increase of frequencies gets smaller as the *OHMS* values increase after 20.

Example 3
Pareto Charts

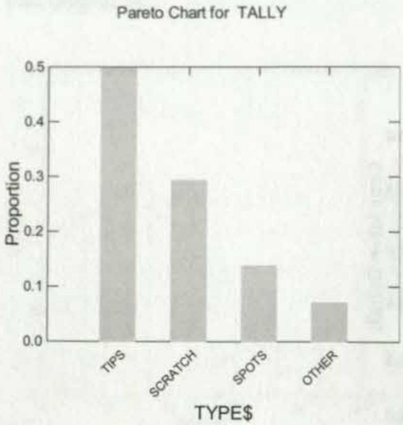
The *CONDENSE* data file contains non-conformance data (defects) for 15 lots of condensers as given by Messina (1987). The variable *LOT$* is lot number, *TYPE* is type of defect, and *TALLY* is the frequency of a particular defect in a particular lot. One thousand condensers were inspected in each lot. Thus, the data are already aggregated by *LOT$* and *TYPE$*. We want a Pareto Chart showing *TALLY* as a proportion of the total defects by *TYPE* of defect.

The input is:

```
QC
   USE CONDENSE
   CATEGORY TYPE$
   PARETO TALLY * TYPE$ / AGG P
```

The output is:

```
Number of Lines of Input Data Read           : 60.000
Number with Missing Data or Zero Weight      :  0.000
Number of Samples to be Plotted              :  4.000
(Only Subgroups Containing Data are Plotted)
```



As you can see from the output, 79% of all the defects tallied were composed of broken tips or scratches on the condensers. In most cases, the data would contain more causes of non-conformities than this (with "other" being broken out into individual types).

Example 4
Box-and-Whisker Plots

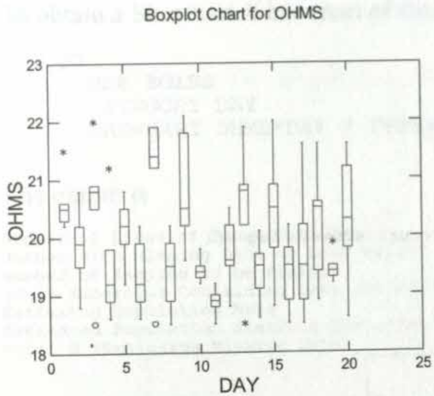
Box plots or "box-and-whisker" plots show a series of box plots for the distribution of a y variable as a function of an x variable that identifies individual samples. For example, the *BOXES* data file contains data showing ohms of electrical resistance for random samples of five computer boxes selected each day for 20 days.

The input is:

```
QC
   USE BOXES
   BOX OHMS * DAY
```

The output is:

```
Number of Lines of Input Data Read          : 100.000
Number with Missing Data or Zero Weight     :   0.000
Number of Samples to be Plotted             : 100.000
(Only Subgroups Containing Data are Plotted)
```



The asterisks indicate that there are one or more data points lying outside the fence for the individual data points. The sample for day 13, for example, has one or more data points below the lower fence for the data; it has no points lying above the upper fence. This makes for a negatively skewed distribution for that sample, which can be seen by the position of the median within the box.

Example 5
Run Chart

A grinding machine used extensively for outer diameter grinding was suffering from tool wear. A number of components are produced from the machine of which one component—adaptor body—was considered for the study.

The data set contains the diameters of 16 components produced over a period of 16 hours one in each hour. The total time period is divided into two periods of eight hours each and the variable *EIGHT* takes value 1 or 2 depending upon the period of its production. Similarly, the variables *FOUR* and *TWO* are constructed. Thus the 'design' is a nested one with *FOUR* nested inside *EIGHT* and *TWO* nested inside *FOUR*.

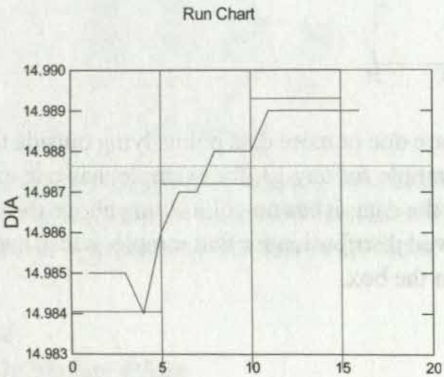
The data for the first 16 components were recorded. The dimension observed in the order of production is given in the data file *ADAPTOR*.

To obtain a Run chart of these data, the input is:

```
QC
USE ADAPTOR
RUNCHART DIA / GROUP = FOUR TREND SHIFT PATTERN
```

The output is:

```
Grouping Variable: FOUR

Mean               : 14.987
Maximum            : 14.989
Minimum            : 14.984
Sample Size        : 16
Range              : 0.005
Standard Deviation : 0.002
Shifts             : Upward
Trends             : None
Patterns           : Some pattern recurs after a lag of1
```



From the output it may seem that there exists some trend, which is missing in the output, but a closer look will show that number of sequential points monotonically increasing (or decreasing) is very few. But the upward movement is very correctly reflected in the shifts of mean. Moreover the patterns in the data set are also reflected in the output.

Example 6
X-bar Chart

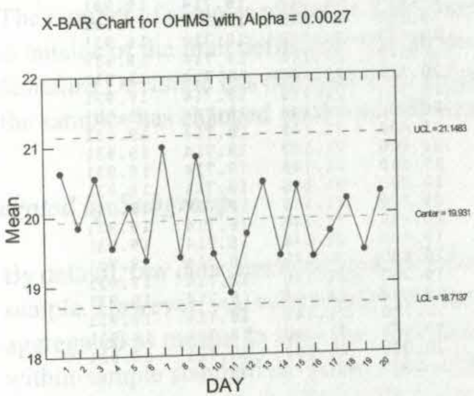
The data set considered here (*BOXES*) has already been described in Example 1. The ohms of electrical resistance in computer boxes were measured for five randomly selected boxes from each of 20 days of production. Thus, each *SAMPLE* contains five observations of resistance in ohms for each of 20 days.

To obtain a Shewhart *X*-bar chart of these data, the input is:

```
QC
   USE BOXES
   CATEGORY DAY
   SHEWHART OHMS*DAY / TYPE=XBAR
```

The output is

```
Number of Lines of Input Data Read              : 100.000
Number with Missing Data or Zero Weight         :   0.000
Number of Samples to be Plotted                 :  20.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       :  19.931
Estimated Population Standard Deviation         :   0.907
Total N (Excluding Missing Data)                :     100
```



The chart shows the mean value of *OHMS* for each *DAY*. The control limits encompass all of the data, so no particular day's mean is out of control, given the default value of alpha that was used.

The output above the chart indicates that the file has 100 cases comprising 20 samples (days). The mean of all 100 cases is shown, as is the pooled within-sample standard deviation. The former is listed as the *Estimated Population Mean*, and the latter is listed as the *Estimated Population Standard Deviation*. It is computed as:

SQR(SS (within samples)/n-k)

where SS (within samples) is the sum of squared deviations around a sample mean pooled across all samples, *n* is the total number of cases for all samples, and *k* is the number of samples. This number is the default value of sigma here and for all Shewhart charts.

Listing Plot Values

To view summary information about each day, set the output length to LONG.

The input is:

```
PLENGTH LONG
SHEWHART OHMS*DAY / TYPE=XBAR
```

The output is:

```
Listing of Chart Data

1  2  3  4  5   6  7  8    Mean     UCL     LCL   Center
-------------------------------------------------------------
                          20.620  21.148  18.714  19.931
                          19.840  21.148  18.714  19.931
                          20.540  21.148  18.714  19.931
                          19.540  21.148  18.714  19.931
                          20.120  21.148  18.714  19.931
                          19.360  21.148  18.714  19.931
            *             20.980  21.148  18.714  19.931
                          19.400  21.148  18.714  19.931
            *             20.840  21.148  18.714  19.931
                          19.440  21.148  18.714  19.931
                          18.880  21.148  18.714  19.931
                          19.720  21.148  18.714  19.931
                          20.460  21.148  18.714  19.931
                          19.400  21.148  18.714  19.931
                          20.400  21.148  18.714  19.931
                          19.380  21.148  18.714  19.931
                          19.740  21.148  18.714  19.931
                          20.200  21.148  18.714  19.931
                          19.460  21.148  18.714  19.931
                          20.300  21.148  18.714  19.931
```

Aggregated by Sample

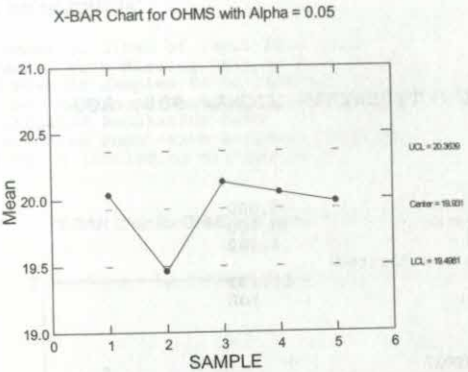
To find out whether any sequential effects are present during the times of day, collapse the data for each sample across days. The *SAMPLE* indicates the sample number within each day. Presumably, these samples were taken in their numerical sequence each day.

The input is:

```
SHEWHART OHMS*SAMPLE / TYPE=XBAR PLIMITS=.025,.975
```

The output is:

```
Number of Lines of Input Data Read              : 100.000
Number with Missing Data or Zero Weight         :   0.000
Number of Samples to be Plotted                 :   5.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       :  19.931
Estimated Population Standard Deviation         :   0.988
Total N (Excluding Missing Data)                :     100
```



The number of samples plotted is 5, the number of samples drawn each day. Sample 2 is outside of the limit defined by this alpha value. The value of *Estimated Population Standard Deviation* has changed from the previous example because the definition of the samples has changed, so the pooled within-sample variance differs.

Aggregated by Subgroup

By default, raw data consist of separate *y* variable values for each individual case in the sample. However, you select Aggregate by subgroup to indicate that input data are aggregated as means. In this case, you must provide a value for Sigma because the within-sample standard deviation cannot then be computed from the input data.

Suppose that the file *MYXBAR.SYD* is created using the output of the data plotted in the previous examples. The resulting data file looks like the following:

SAMPLE	MEAN	CENTER	UCL	LCL	N
1.000	20.035	19.931	20.594	19.268	20.000
2.000	19.460	19.931	20.594	19.268	20.000
3.000	20.130	19.931	20.594	19.268	20.000
4.000	20.050	19.931	20.594	19.268	20.000
5.000	19.980	19.931	20.594	19.268	20.000

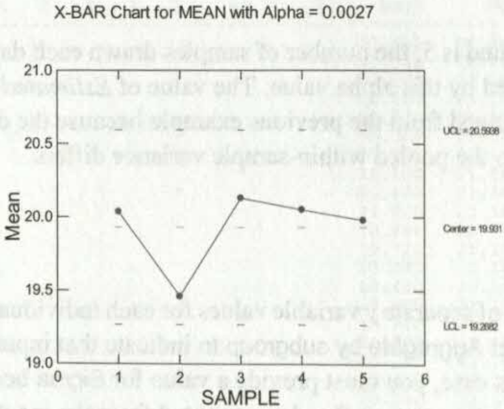
To plot this aggregated data in an *X*-bar chart, you must specify the within-sample standard deviation (SIGMA) for whatever sample identifier you use. The output from the previous raw data run displays this value: 0.988. To specify that variable *N* contains the sample sizes, FREQUENCY is used.

The input is:

```
USE MYXBAR
FREQUENCY N
SHEWHART MEAN*SAMPLE / TYPE=XBAR SIGMA=.988, AGG
```

The output is:

```
Number of Lines of Input Data Read              :  5.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 :  5.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       : 19.931
Total N (Excluding Missing Data)                :    100
```



The estimated population standard deviation is not reported. It cannot be estimated from the data because there is no within-sample variance (each case represents a separate sample). That is why Sigma must be specified.

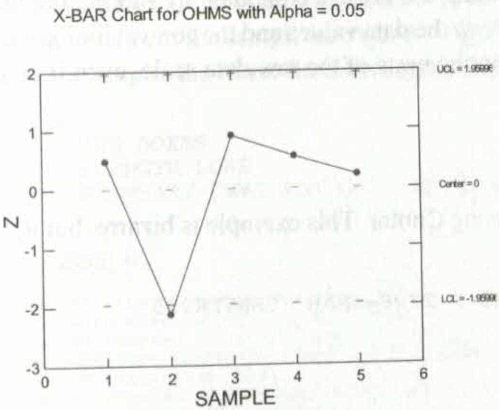
Standardizing the Data

Now let us look at the chart for the five samples using the Standard deviation chart units (*Z*) option to standardize the data. We reopen the *BOXES* data file.

```
USE BOXES
SHEWHART OHMS*SAMPLE / TYPE=XBAR  PLIMITS=.025,.975 Z
```

The output is:

```
Number of Lines of Input Data Read                  : 100.000
Number with Missing Data or Zero Weight             :   0.000
Number of Samples to be Plotted                     :   5.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                           :  19.931
Estimated Population Standard Deviation             :   0.988
Total N (Excluding Missing Data)                    :     100
```

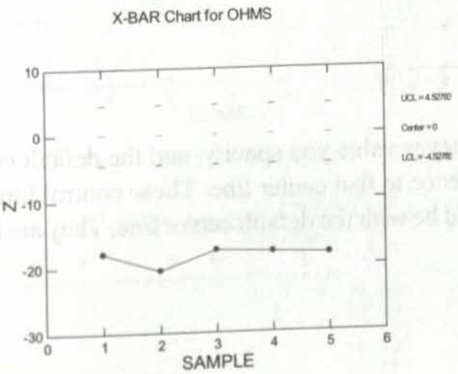


Control Limits

Here is a somewhat bizarre example intended to make a point. It is a *Z* chart, but this time with the center line and control limits specified in advance.

```
SHEWHART OHMS*SAMPLE / TYPE=XBAR CENTER=24 LIMITS = 23, 25, Z
```

The output is:



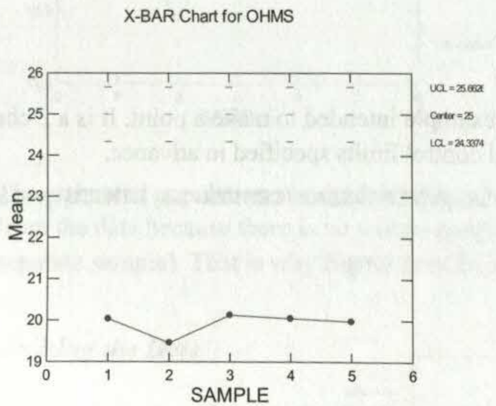
We specify extreme values for Center and Limits to set the limits apart from the data. This shows what had happened in the transformation from the original means to Z values. Notice that all of the data are converted to Z values, using the center line of 24 as the origin for Z. The lower control limit is computed by subtracting 24 from 23 and dividing by the standard error of the mean for each sample. Similarly, the upper limit is computed by subtracting 24 from 25 and dividing by the standard error for each sample. The point is that Z values are always computed by subtracting the specified value of the center line from *both* the data values and the control limits. Further, Center and Limits are always stated in the units of the raw data scale, even if Z is requested.

Specifying the Center

Let us look at the effects of using Center. This example is bizarre, but it, too, makes a point.

```
SHEWHART OHMS*SAMPLE / TYPE=XBAR  CENTER=25
```

The output is:



The center line is placed at whatever value you specify, and the default control limits are always computed with reference to that center line. These control limits are the same distance apart as they would be with the default center line. They are just centered at a different point on the chart.

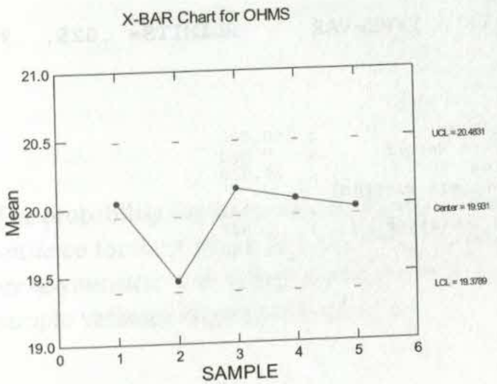
Sigma Limits

This example displays sigma limits (SLIMITS), or units of the standard deviation of the sampling distribution for each sample. The default sigma limits are set at –3 and +3. In other words, the limits that you get are three standard deviations of the sampling distribution on either side of the center line (Notice that the standard deviation of the sampling distribution is not sigma or its default value. Rather, it is sigma divided by the square root of the sample size when we deal with a sampling distribution of means). In this example, the lower limit and upper limit are set to produce 2.5-sigma limits.

```
QC
   USE BOXES
   PLENGTH LONG
   SHEWHART OHMS*SAMPLE / TYPE=XBAR SLIMITS=-2.5, 2.5
```

The output is:

```
Number of Lines of Input Data Read              : 100.000
Number with Missing Data or Zero Weight         :   0.000
Number of Samples to be Plotted                 :   5.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       :  19.931
Estimated Population Standard Deviation         :   0.988
Total N (Excluding Missing Data)                :     100
```




```
Listing of Chart Data
1  2  3  4  5  6  7  8       Mean      UCL      LCL   Center
------------------------------------------------------------
                           20.035   20.483   19.379   19.931
                           19.460   20.483   19.379   19.931
                           20.130   20.483   19.379   19.931
                           20.050   20.483   19.379   19.931
                           19.980   20.483   19.379   19.931
```

The *Estimated Population Standard Deviation* is 0.988, and the sample size is 20 for all samples. Thus, the standard deviation of the sampling distribution of the mean is 0.988/SQR(20), which equals 0.221. Multiplying this by 2.5 and then adding and subtracting it from the value of center yields 20.483 and 19.379, the respective values of the limits in the output. The limits are the same for each sample, but only because the sample size remains constant. If it did not, we would get limits that vary from sample to sample. Also notice that the limits would differ if, in addition, we use Sigma. Then the value supplied for Sigma is used to calculate the limits, rather than the empirical value of the *Estimated Population Standard Deviation* in the output.

Example 7
Variance Chart

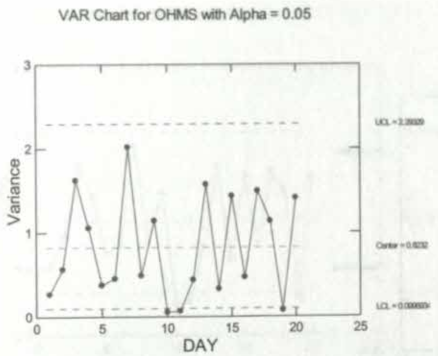
This example produces a variance chart for the variable *OHMS* (from the *BOXES* data file).

The input is:

```
QC
   USE BOXES
   PLENGTH LONG
   SHEWHART OHMS*DAY  / TYPE=VAR     PLIMITS= .025, .975
```

The output is:

```
Number of Lines of Input Data Read              : 100.000
Number with Missing Data or Zero Weight         :   0.000
Number of Samples to be Plotted                 :  20.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       :  19.931
Estimated Population Standard Deviation         :   0.907
Total N (Excluding Missing Data)                :     100
```



Listing of Chart Data

1	2	3	4	5	6	7	8	Variance	UCL	LCL	Center
								0.267	2.293	0.100	0.823
								0.563	2.293	0.100	0.823
								1.623	2.293	0.100	0.823
								1.058	2.293	0.100	0.823
								0.377	2.293	0.100	0.823
								0.453	2.293	0.100	0.823
								2.017	2.293	0.100	0.823
								0.490	2.293	0.100	0.823
								1.143	2.293	0.100	0.823
								0.053	2.293	0.100	0.823
								0.067	2.293	0.100	0.823
								0.437	2.293	0.100	0.823
								1.573	2.293	0.100	0.823
								0.330	2.293	0.100	0.823
								1.435	2.293	0.100	0.823
								0.467	2.293	0.100	0.823
								1.493	2.293	0.100	0.823
								1.135	2.293	0.100	0.823
								0.073	2.293	0.100	0.823
								1.410	2.293	0.100	0.823

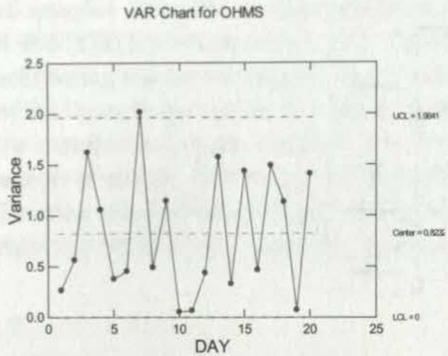
The probability limits are set at 0.025 and 0.975, producing an alpha of 0.05. The variance for days 10, 11 and 19 fall below the lower control limit. The control limits are asymmetric with regard to the center line because the sampling distribution of the sample variance is positively skewed.

Sigma Limits

This example uses sigma limits instead of probability limits. To compare the results with the previous example, we use 1.96 sigma limits because 1.96 is the standard normal deviate that produces a two-tailed alpha of 0.05.

```
SHEWHART OHMS*DAY / TYPE=VAR   SLIMITS=-1.96,1.96
```

The output is:



None of the small sample variances are below the LCL. They cannot be, because the LCL is 0. (Subtracting 1.96 standard deviations from its mean resulted in a value less than 0, so SYSTAT set the LCL to 0.) Compare this with the previous example, in which three of the samples showed variances below the LCL. Also, note that sample 7 has a variance that exceeds the UCL in the present example, but does not in the previous example.

Example 8
s chart

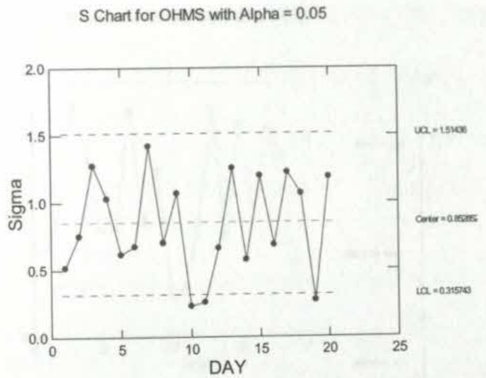
The *s* chart plots sample standard deviation as a function of a sample identifier.

The input is:

```
QC
  USE BOXES
  SHEWHART OHMS*DAY / TYPE=S  PLIMITS=.025,.975
```

The output is:

```
Number of Lines of Input Data Read               : 100.000
Number with Missing Data or Zero Weight          :   0.000
Number of Samples to be Plotted                  :  20.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                        :  19.931
Estimated Population Standard Deviation          :   0.907
Total N (Excluding Missing Data)                 :   100
```



As you can see, samples for days 10, 11, and 19 fall below the lower control limit, just as they did for the comparable variance chart. This is expected because sample standard deviations are just the square roots of sample variances. The control limits are less asymmetric in this chart because of this square-root function.

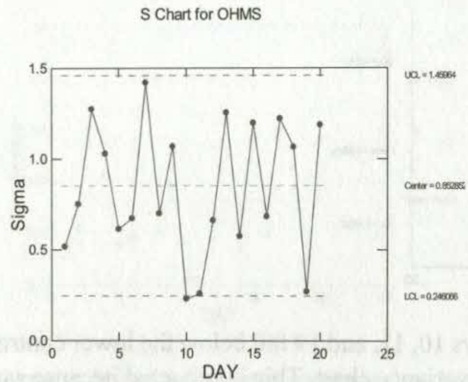
Sigma Limits

You can plot *s* chart data with sigma limits, rather than probability limits, and compare the results. The previous chart had probability limits set at 0.025, 0.975 (which sets alpha to 0.05). The comparable limits for a sigma-limits chart are –1.96 and +1.96 standard deviations.

The modified input line is:

```
SHEWHART OHMS*DAY / TYPE=S  SLIMITS=-1.96,1.96
```

The output is:



The values for days 11 and 19 are not outside the lower control limit as they were when we used probability limits in the other *s* chart example.

Example 9
R Chart

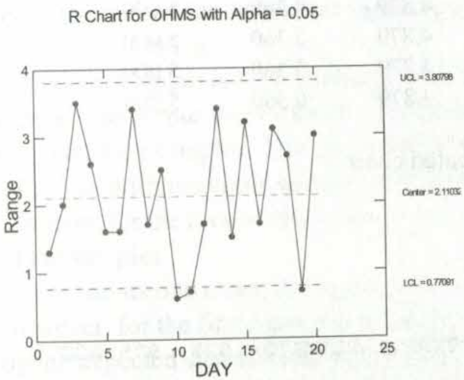
Here we produce an *R* chart for the *BOXES* data.

The input is:

```
QC
   USE BOXES
   SHEWHART OHMS*DAY / TYPE=R  PLIMITS=.025,.975
```

The output is:

```
Number of Lines of Input Data Read              : 100.000
Number with Missing Data or Zero Weight         :   0.000
Number of Samples to be Plotted                 :  20.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       :  19.931
Estimated Population Standard Deviation         :   0.907
Total N (Excluding Missing Data)                :     100
```



As for the variance and *s* chart examples, the samples for days 10, 11, and 19 are slightly out side control limits when alpha=0.05. Even though ranges are computed, SYSTAT still uses all of the data to produce values for the grand mean and pooled within-sample standard deviation and so these values are the same as those listed for earlier examples.

Aggregated Data

It seems a waste of data to run an R chart using a file that contains all the data. The more likely use of this chart is when you have the range for each sample. All you need is a sample identifier, the range for each sample, and the sample size. Then use Aggregate by subgroup to analyze the data.

For example, you can create a data set named *MYR* by preceding the SHEWHART command for the raw data with a SAVE command.

The input is:

```
SAVE MYR
SHEWHART OHMS*DAY/TYPE=R
```

MYR contains the following data:

DAY	RANGE	CENTER	UCL	LCL	N
1.000	1.300	2.110	4.879	0.360	5.000
2.000	2.000	2.110	4.879	0.360	5.000
3.000	3.500	2.110	4.879	0.360	5.000
4.000	2.600	2.110	4.879	0.360	5.000
...	...	...	...	...	...

17.000	3.100	2.110	4.879	0.360	5.000
18.000	2.700	2.110	4.879	0.360	5.000
19.000	0.700	2.110	4.879	0.360	5.000
20.000	3.000	2.110	4.879	0.360	5.000

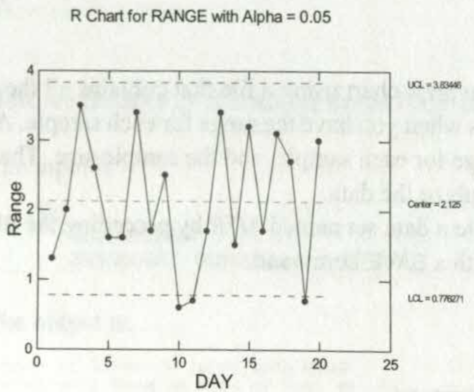
Use this file to create the aggregated chart.

The input is:

```
QC
   USE MYR
   FREQUENCY N
   SHEWHART RANGE*DAY/TYPE=R PLIMITS=.025,.975 AGG
```

The output is:

```
Number of Lines of Input Data Read              : 20.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 : 20.000
(Only Subgroups Containing Data are Plotted)
Mean of Sample Ranges                           :  2.125
Estimated Population Standard Deviation         :  0.914
Total N (Excluding Missing Data)                :    100
```



As before, the ranges for days 10, 11, and 19 are out of control limits. Notice the differences in the header information for this chart: it shows the mean of the input "*RANGES*" rather than the mean of all data points. This mean is the center line of the chart. Also, notice that the estimated population standard deviation differs somewhat from the previous example, because the within-sample data points are not in the input file. SYSTAT estimates the population sigma value from the mean of the sample ranges, producing a value of 0.914 instead of 0.907. The value 0.907 is more reliable

since the range is an inefficient estimate of sigma. But if ranges are all you have, it is better than nothing.

If output length is set to LONG for the *R* chart examples (set Options or PLENGTH LONG), the output shows that the center line and control limit values differ slightly between the examples. The sigma value for the first chart is calculated from the empirical within-sample standard deviation (0.907) based on all the individual cases of raw data; for the second chart, sigma is an estimate based on the average of the ranges of the samples.

For the second chart, the center line is exactly the mean of the sample ranges. However, for the first chart, the center line is computed as the sigma estimate divided by the expected standardized range. One could argue that the center lines for both charts should be just the mean of the sample ranges. Indeed, that is the usual way of estimating the expected value from a sampling distribution. It seems more theoretically sound to use the sigma estimate based on all the data, when available, to estimate the expected value of the range, rather than the mean of the sample ranges, because the mean range is not reflective of the raw data.

Example 10
np Chart

For this chart, it is unlikely that you will want to enter raw data, which would be a sequence of 0's and 1's for each sample. A less tedious way is to use FREQUENCY. For example, suppose that five machines each produce a product. You sample a sequence of 100 items from each machine at approximately the same time. Each item is classified as either a 0 if the product conforms to specifications or a 1 if it does not. In the following data set, the variable *N* contains the counts of conforming (*RESULT*=1) and non-conforming units (*RESULT*=0) for each machine:

MACHINE$	RESULT	N
1	1	20
1	0	80
2	1	10
2	0	90
3	1	30
3	0	70
4	1	40
4	0	60
5	1	0
5	0	100

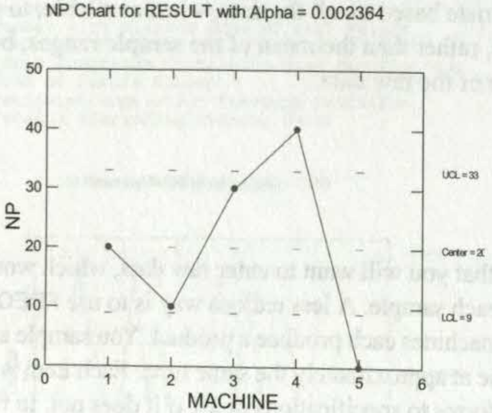
You can analyze this data set as follows.

The input is:

```
QC
    USE MACHINE
    FREQUENCY N
    SHEWHART RESULT*MACHINE / TYPE=NP
```

The output is:

```
Number of Lines of Input Data Read                 : 10.000
Number with Missing Data or Zero Weight             :  1.000
Number of Samples to be Plotted                    :  5.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Proportion(p) value           :  0.200
Total N (Excluding Missing Data)                   :    500
```



The output reports that one line of input data either had missing data or zero weight because machine 5 had a value of 0 for *N*. When SYSTAT encounters a frequency value of 0, the case is ignored. (Here we simply want to multiply the value of *RESULT*, which is 1, by *N*, which is 0. If we do, and enter into the total, it has no effect. Ignoring this data produces exactly the same result.)

Notice that the estimated population *p* value is reported rather than the estimated population mean, as in some of the previous charts. The overall proportion of 1's in the data set is, in fact, the mean of all the data, and therefore, it is an estimate of the population mean for a Bernoulli distribution. It is simply labeled differently here for clarity.

In the graph, a data value falls at the same location as the center line. Thus, there are two pieces of information at the same point on the chart. The data values for machines

4 and 5 are outside the control limits. Machine 4 produces too many nonconforming items. Machine 5, on the other hand, produced none, making this sample a candidate for faulty inspection procedures. Finally, notice that the control limits have integer values. The binomial is a discrete distribution and so all values on the chart except the center line are limited to integer values.

Aggregated Data

Now let us look at an example using aggregated data. The sample data file *JUICE* contains data from Montgomery (2001) on the number of defective orange juice cans found in each of 24 samples of 50 juice cans. The data were collected on each of three shifts with eight samples taken for each shift. The data set has the following structure:

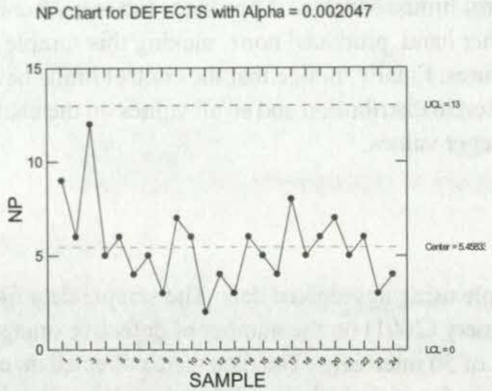
SAMPLE	SHIFT	TIME	DEFECTS	SIZE
1.000	1	1	9.000	50.000
2.000	1	2	6.000	50.000
3.000	1	3	12.000	50.000
4.000	1	4	5.000	50.000
5.000	1	5	6.000	50.000
…	…	…	…	…
19.000	3	3	6.000	50.000
20.000	3	4	7.000	50.000
21.000	3	5	5.000	50.000
22.000	3	6	6.000	50.000
23.000	3	7	3.000	50.000
24.000	3	8	4.000	50.000

The input is:

```
USE JUICE
FREQUENCY SIZE
CATEGORY SAMPLE
SHEWHART DEFECTS*SAMPLE / TYPE=NP AGG
```

The output is:

```
Number of Lines of Input Data Read              : 24.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 : 24.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Proportion(p) value        :  0.109
Total N (Excluding Missing Data)                :   1200
```



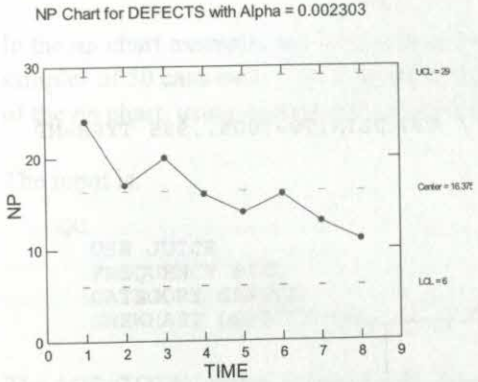
The default probability limit values are used. However, the reported alpha in the title of the graph is 0.00204 — which is not the default alpha. The reported alpha is as close to the desired (default) alpha as possible without exceeding it. This discrepancy is due to the discrete value of the binomial distribution.

Pooling Results

Now let us rerun this feature with the aggregated data example but plot by *TIME* instead of by *SAMPLE*. This illustrates the fact that SYSTAT "pools" binomial results over subgroups within a sample. Assuming the *JUICE* file is still in use (and FREQUENCY SIZE is still in effect), the additional input is:

```
SHEWHART DEFECTS*TIME/TYPE=NP AGG
```

The output is:



A downward trend in defects across time is evident, indicating that the process is getting more and more in control as time wears on within shifts. Note that we could also plot these data by *SHIFT*, pooling across *TIME* within *SHIFT*.

Varying Sample Sizes

Let us look at the data presented by Montgomery (2001) that helps understand an important aspect of *np* charts. In this example, all units of a personal computer produced on each of 10 successive days were inspected. Let *UNITS* be the number of computers inspected each day, and let *NONCON* be the number of non-conforming units found. The data are in the *COMPUTER* data file.

The data set is as follows:

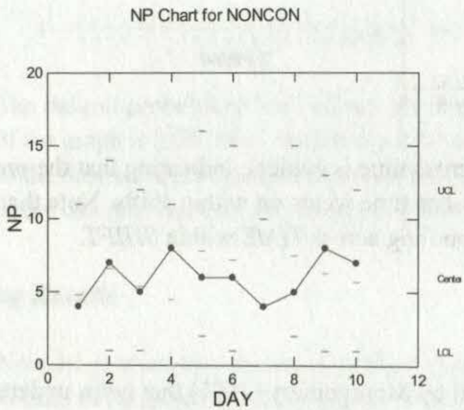
DAY	UNITS	NONCON
1	80	4
2	110	7
3	90	5
4	75	8
5	130	6
6	120	6
7	70	4
8	125	5
9	105	8
10	95	7

We analyze these aggregated data using FREQUENCY to indicate that *UNITS* contains sample sizes for each sample.

The input is:

```
USE COMPUTER
FREQUENCY UNITS
SHEWHART NONCON*DAY / AGG PLIMITS=.005,.995 TYPE=NP
```

The output is:



The center line and the control limits are extremely varied because the number of units inspected each day varies. The control limits and the center line are both joint functions of the overall proportion of non-conforming units and the sample size. While the chart is not pleasing to the eye, it is, nevertheless, interpretable. One must look carefully at the individual center values and limits for each sample to see where the number of non-conforming units falls relative to them. This unaesthetic result is a hazard due to 100% inspection when the number of units varies among the subgroups used to define samples (in this case, days).

There are three ways to make this control chart more visually appealing. One is to run a *p* chart instead of an *np* chart, which at least makes the center line a constant for all samples. This is exemplified for the same data set in the discussion of *p* charts below. A second way is to use the standard deviation chart units (*Z*) option to convert to *Z* values, which produces a constant center line. Third, we could use the Average subgroup size (AVGN) option, which artificially treats the mean of all sample sizes as the sample size for each sample. This third option is not desirable here because the sample sizes differ widely among days.

Example 11
p Chart

In the *np* chart example, we analyze data for the number of defective juice cans in 24 samples of 50 cans each. This example reanalyzes these data using the *p* chart instead of the *np* chart, using sigma limits instead of probability limits.

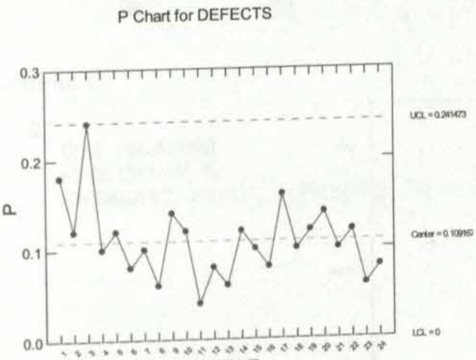
The input is:

```
QC
   USE JUICE
   FREQUENCY SIZE
   CATEGORY SAMPLE
   SHEWHART DEFECTS*SAMPLE/ TYPE =P AGG=TOTAL SLIMITS = -3, 3
```

The AGG=TOTAL option is used (in the dialog box, select Indicate subgroup size) because the data are total counts rather than proportions; this option allows you to enter total counts as aggregated data.

The output is:

```
Number of Lines of Input Data Read                : 24.000
Number with Missing Data or Zero Weight           :  0.000
Number of Samples to be Plotted                   : 24.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Proportion(p) value          :  0.109
Total N (Excluding Missing Data)                  :  1200
```



The upper limit for this chart is at 0.24. This corresponds to 12 defects in a sample of 50. The *np* chart for the same data used probability limits, and the upper limit was 13 instead of 12. Thus, the probability limits and the sigma limits yield approximately the

same results in this case. However, a difference of 1 might possibly be important to your quality program.

As with the *np* chart, you could plot the data in the *JUICE* file by *SHIFT* or by *TIME*.

Aggregated Data

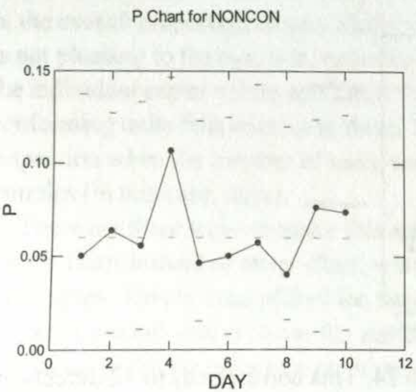
Another *np* chart example involved the number of non-conforming computer units produced each day using 100% inspection. Recall that this produced an unaesthetic chart because of the unequal number of units inspected each day. Let us try a *p* chart for the same data, using FREQUENCY to designate sample size and AGG=TOTAL because aggregate data are counts rather than proportions.

The input is:

```
USE COMPUTER
FREQUENCY UNITS
SHEWHART NONCON*DAY / TYPE=P AGG=TOTAL PLIMITS=.005,.995
```

The output is:

```
Number of Lines of Input Data Read              : 10.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 : 10.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Proportion(p) value        :  0.060
Total N (Excluding Missing Data)                :  1000
```



This chart is clearer than its *np* counterpart because the center line is not variable. Control limits still vary from sample to sample because they are based on sample sizes

that vary. All of the comments regarding the binomial *np* chart also apply here because *p* is just *np*/*n*.

Example 12
c Chart

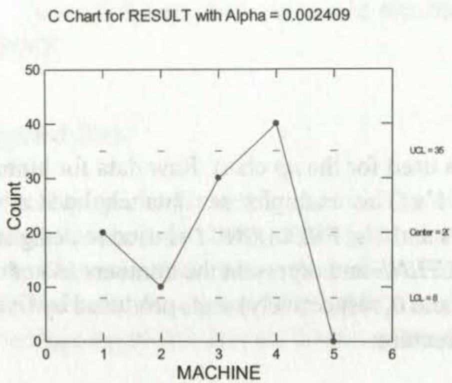
This example uses raw input data used for the *np* chart. Raw data for binomial and Poisson charts consist of 0's and 1's (This example uses data tabulated as counts to avoid entering a long string of 0's and 1's; FREQUENCY is used to designate sample size). The data are in the file *MACHINE* and represent the numbers (*N*) of conforming and nonconforming (*RESULT*=1 and 0, respectively) units produced by five machines. The data set has the following structure:

MACHINE$	RESULT	N
A	1	20
A	0	80
B	1	10
B	0	90
C	1	30
C	0	70
D	1	40
D	0	60
E	1	0
E	0	100

The input is:

```
QC
   USE MACHINE
   FREQUENCY N
   SHEWHART RESULT*MACHINE/TYPE=C
```

The output is:



Machines 4 and 5 show results outside the control limits, just as they did for the *np* chart. In the *np* chart, the control limits were at 9 and 33; in this chart, they are at 8 and 35 because the Poisson distribution is used to approximate the binomial. The limits are integer values because the Poisson distribution is discrete.

Aggregated Data

Now let us use aggregated data from Montgomery (2001) on the number of non-conformities found in 26 successive samples of 100 circuit boards. For convenience, the sample unit (or inspection unit) is defined as 100 boards. That is, although each sample contains 100 boards, each sample is considered a sample of size 1 from a Poisson distribution. The data are in the *BOARDS* file with the following structure:

SAMPLE	DEFECTS
1.000	21.000
2.000	24.000
3.000	16.000
…	…
24.000	19.000
25.000	17.000
26.000	15.000

SAMPLE is the identifier and *DEFECTS* is a total count of the number of defects in each group of 100 boards.

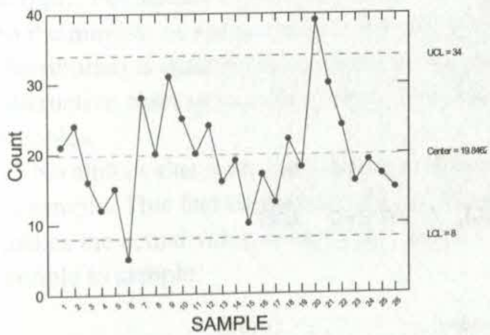
The input is:

```
USE BOARDS
CATEGORY SAMPLE
SHEWHART DEFECTS*SAMPLE/TYPE=C AGG
```

Notice that AGG (Aggregate by subgroup) is used because the data are total counts for each sample. Each of the 26 samples constitutes exactly one sample unit. FREQUENCY is not needed because the default frequency for each sample is 1. The default values for probability limits are used, resulting in alpha = 0.0027.

The output is:

```
Number of Lines of Input Data Read              : 26.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 : 26.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Count/Sample-Unit          : 19.846
Total Units (Excluding Missing Data)            : 26.000
```



The estimated population count of defects per sample unit is 19.846 (There are 516 defects overall in the 26 sample units). This becomes the center line for the c chart. The output variable *COUNT* contains the same data as the input variable *DEFECTS*.

Montgomery's calculations of the upper control limit for these data lead to a value of 33.22 using the traditional 3-sigma limit approach. The lower limit he obtains is 6.48. These are approximately the same as the limits found here using the actual Poisson distribution probability limits. The SYSTAT probability limits are more accurate. If we had used sigma limits for the *c* chart, we would find the same limits that Montgomery did.

Sample Units

Montgomery's (2001) data count the occurrences of non-conformities in each of 10 rolls of dyed cloth. The rolls are not the same size in square meters. Thus, the sample unit is defined as 50 square meters of cloth, and roll sizes are expressed in these units. The first roll contains 500 square meters of cloth and the second contains 400 square meters. Thus, the sample units for these rolls are 500/50 = 10 and 400/50 = 8, respectively. The remainder of the sample units are calculated similarly. These data are stored in the *CLOTH* file with the following structure:

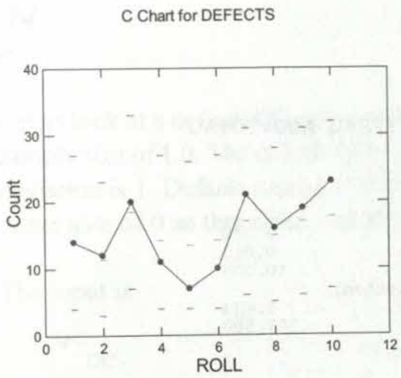
ROLL	DEFECTS	UNITS
1	14	10.0
2	12	8.0
3	20	13.0
4	11	10.0
5	7	9.5
6	10	10.0
7	21	12.0
8	16	10.5
9	19	12.0
10	23	12.5

The input is:

```
USE CLOTH
WEIGHT UNITS
SHEWHART DEFECTS*ROLL / TYPE=C  AGG
```

The output is:

```
Number of Lines of Input Data Read              :  10.000
Number with Missing Data or Zero Weight         :   0.000
Number of Samples to be Plotted                 :  10.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Count/Sample-Unit          :   1.423
Total Units (Excluding Missing Data)            : 107.500
```



Here, as with the binomial example, we have an unclear chart because the number of units differs from sample to sample. A total of 153 non-conformities (defects) occur in the 10 rolls. The total sample units is 107.50, so the mean number of non-conformities per sample unit was 153/107.5 = 1.423. The center line for each sample is 1.423 times the number of sample units for that sample. Thus, the center line varies from sample to sample. For similar reasons, the control limits vary, their widths being directly related to the number of sample units involved (Recall that the variance of a Poisson distribution is directly proportional to the mean). The *u* chart example uses an alternative chart to reanalyze these data, making the center line a constant for all samples.

No alpha value is reported, because the number of sample units differs from sample to sample. This fact combined with the discrete nature of the Poisson distribution makes the actual value of alpha (as opposed to the requested value) different from sample to sample.

Example 13
u Chart

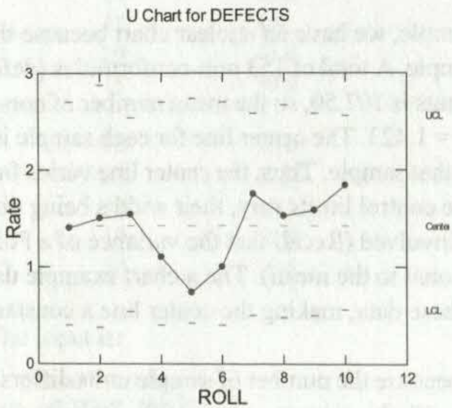
This example uses the *CLOTH* data file, described in the *c* chart examples. Because the data are total counts of defects per roll (rather than rate by roll), AGG=TOTAL (Indicate subgroup size) is used.

The input is:

```
QC
   USE CLOTH
   WEIGHT UNITS
   SHEWHART DEFECTS*ROLL/TYPE=U AGG=TOTAL
```

The output is:

```
Number of Lines of Input Data Read            :  10.000
Number with Missing Data or Zero Weight       :   0.000
Number of Samples to be Plotted               :  10.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Count/Sample-Unit        :   1.423
Total Units (Excluding Missing Data)          : 107.500
```



The output information in the header is the same as that in the *c* chart plotted for the same data. No value of alpha is reported for this *u* chart because the number of sample units differs from sample to sample. This fact combined with the discrete nature of the Poisson distribution makes the actual value of alpha (as opposed to a requested value) different from sample to sample. Thus, no single value suffices.

The data on this chart are the corresponding data for the *c* chart divided by the number of sample units for each sample. The center line is constant for this chart, making it easier to read. The limits still differ by sample because the number of sample units varies from sample to sample.

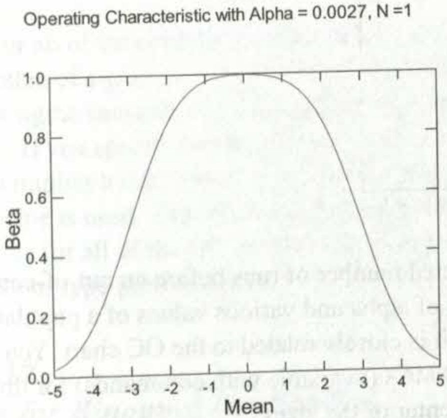
Example 14
OC Curve

Let us look at a default OC curve with a standard normal distribution of means using a sample size of 1.0. The default null hypothesis mean is then 0, and the default standard deviation is 1. Default critical values are set at three standard errors of the mean on either side of 0 so that alpha = 0.0027.

The input is:

```
QC
   OC
```

The output is:



Here you see the probability of a Type II error (beta) plotted as a function of possible true values of the population mean. The value of beta when the population mean is 0 is simply 1 minus alpha. Because the default critical values were at –3 and +3, this value of beta is 1.0000 – 0.0027 = 0.9973. Thus, the probability of accepting the null hypothesis when it is in fact true is 0.9973.

Example 15
ARL Curve

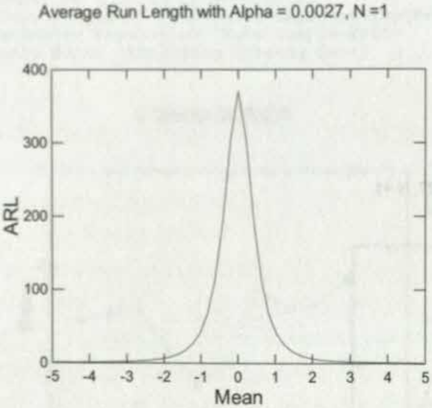
Let us look at a default ARL curve for a standard normal distribution of means using a sample size of 1.0. The default null hypothesis mean is then 0, and the default standard

deviation is 1. Default critical values are set at three standard errors of the mean on either side of 0 so that alpha = 0.0027.

The input is:

```
QC
   ARL
```

The output is:



The vertical axis shows the expected number of runs before an out-of-control signal would appear for the stated value of alpha and various values of a population mean. ARL is defined as 1/(1–beta), so it is closely related to the OC chart. You might want to set tighter values of XMIN and XMAX (available with commands) for this chart to improve the resolution near the center of the chart.

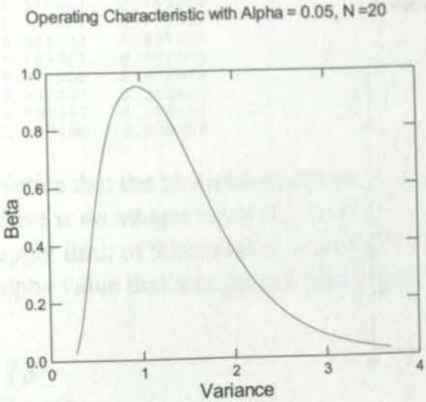
Example 16
OC Curve for Variances

Now let us look at an OC plot of values of the population variance. No data file is used. Specify probability limits (PLIMITS) of 0.025 and 0.975, and a sample size of 20. As no value is specified for Center, its default value is computed as 1.

The input is:

```
QC
   OC / TYPE=VAR  N=20 PLIMITS=.025,.975
```

The output is:



For all of the continuous distributions used in the OC and ARL commands, the default value of sigma is 1. Since the null hypothesis (center) value for a variance type of OC is sigma squared, the center value will also be 1 by default when Sigma is not specified.

If you specify Center, that is the value used as the null hypothesis value for the plot. It implies a sigma value. Similarly, if you specify a value for Sigma, the implied center value is used. Therefore you will get an error message if you specify both. This holds true for all of the OC or ARL commands for continuous distributions except for the *X*-bar type (normal) plot.

Example 17
OC Curve for Binomial Distribution

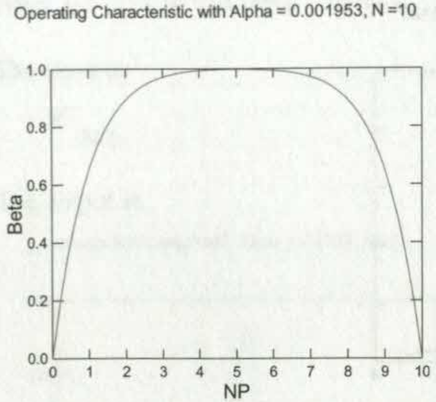
Let us consider a discrete distribution—the binomial. SYSTAT requires that you specify Center for this distribution because there is no sensible default. You cannot specify Sigma because Center determines Sigma. This is true for OC or ARL curves for the *np, p, c*, or *u* types.

Let us plot an OC curve for the binomial distribution with 10 trials and a population probability of 0.5. Note that the center value is then 10 times 0.5. To display the chart values, PLENGTH LONG is specified.

The input is:

```
QC
   PLENGTH LONG
   FORMAT 6
   OC / TYPE=NP  N=10  CENTER=5  PLIMITS=.001,.999
```

The output is:



Operating Characteristic Data

NP	Beta
0.000000	0.000000
0.208333	0.189849
0.416667	0.346620
0.625000	0.475540
0.833333	0.581096
1.041667	0.667132
1.250000	0.736924
1.458333	0.793259
1.666667	0.838494
1.875000	0.874618
2.083333	0.903300
2.291667	0.925936
2.500000	0.943686
2.708333	0.957510
2.916667	0.968199
3.125000	0.976401
3.333333	0.982642
3.541667	0.987345
3.750000	0.990850
3.958333	0.993426
4.166667	0.995280
4.375000	0.996572
4.583333	0.997417
4.791667	0.997893
5.000000	0.998047
5.208333	0.997893
5.416667	0.997417
5.625000	0.996572
5.833333	0.995280
6.041667	0.993426
6.250000	0.990850
6.458333	0.987345
6.666667	0.982642
6.875000	0.976401
7.083333	0.968199
7.291667	0.957510
7.500000	0.943686
7.708333	0.925936
7.916667	0.903300

```
 8.125000    0.874618
 8.333333    0.838494
 8.541667    0.793259
 8.750000    0.736924
 8.958333    0.667132
 9.166667    0.581096
 9.375000    0.475540
 9.583333    0.346620
 9.791667    0.189849
10.000000    0.000000
```

Notice that the binomial distribution is discrete, and, for this binomial distribution, there is no integer value above which lies 0.002 of the distribution. SYSTAT used an upper limit of 9 because decreasing the upper limit to 8 would have created an actual alpha value that was greater than an alpha value of 0.002.

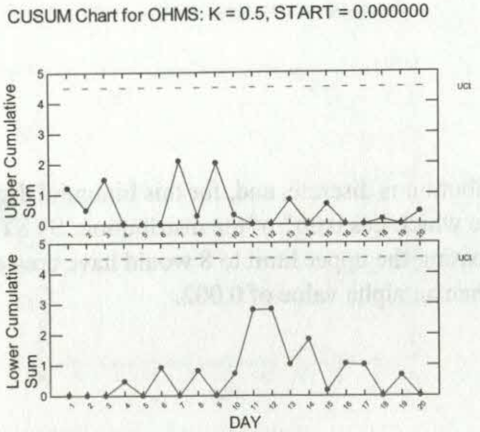
Example 18
Cusum Charts

The following Cusum chart example uses the *BOXES* data file. The *Z* control limit is set to 2.0 to ensure that a few *Z* values will exceed ZCL, just to show you what happens when they do.

The input is:

```
QC
   USE BOXES
   PLENGTH LONG
   CATEGORY DAY
   CUSUM OHMS*DAY / ZCL=2.0  H=4.5
```

The output is:




```
Out-of-control samples are marked with * in the first column.
K : 0.500   H : 4.500   Absolute Z-Limit : 2.000

Listing of Chart Data
```

	DAY	CUSUM_HI	CUSUM_LO	Z	Mean	N
	1.000	1.198	0.000	1.698	20.620	5.000
	2.000	0.474	0.000	-0.224	19.840	5.000
	3.000	1.475	0.000	1.501	20.540	5.000
	4.000	0.011	0.464	-0.964	19.540	5.000
	5.000	0.000	0.000	0.466	20.120	5.000
	6.000	0.000	0.907	-1.407	19.360	5.000
*	7.000	2.085	0.000	2.585	20.980	5.000
	8.000	0.277	0.809	-1.309	19.400	5.000
*	9.000	2.017	0.000	2.240	20.840	5.000
	10.000	0.307	0.710	-1.210	19.440	5.000
*	11.000	0.000	2.800	-2.590	18.880	5.000
	12.000	0.000	2.820	-0.520	19.720	5.000
	13.000	0.804	1.017	1.304	20.460	5.000
	14.000	0.000	1.825	-1.309	19.400	5.000
	15.000	0.656	0.169	1.156	20.400	5.000
	16.000	0.000	1.027	-1.358	19.380	5.000
	17.000	0.000	0.998	-0.471	19.740	5.000
	18.000	0.163	0.000	0.663	20.200	5.000
	19.000	0.000	0.661	-1.161	19.460	5.000
	20.000	0.409	0.000	0.909	20.300	5.000

Both charts contain only non-negative numbers because of the method of calculation, so only an upper control limit, *H*, is required for either chart. You can see from the charts and the tabular output that the cumulative sum did not approach out of control. However, the asterisks in the tabular output indicate that three *Z* values were greater than the absolute value set with the ZCL option. Had any values of the cumulative sum exceeded *H*, those values also would have been flagged with asterisks.

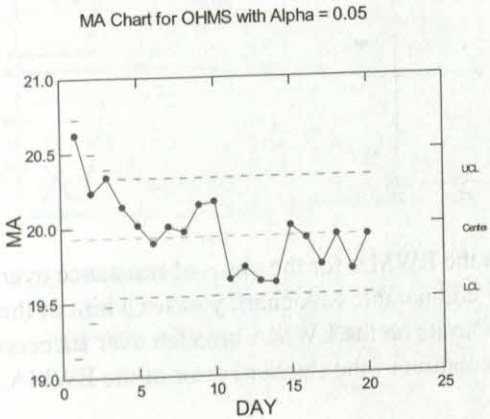
Example 19
Moving Average Chart

Moving average (MA) charts show the unweighted moving average of a *y* variable as a function of an *x* variable that identifies each sample. For example, to produce an MA chart for the *BOXES* data file, the input is:

```
QC
   USE BOXES
   MA OHMS*DAY / PLIMITS=.025, .975 WIDTH=4 YMIN=19.0,
                       YMAX=21.0
```

The output is:

```
Number of Lines of Input Data Read                    : 100.000
Number with Missing Data or Zero Weight               :   0.000
Number of Samples to be Plotted                       :  20.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                             :  19.931
Estimated Population Standard Deviation               :   0.907
Total N (Excluding Missing Data)                      :     100
```



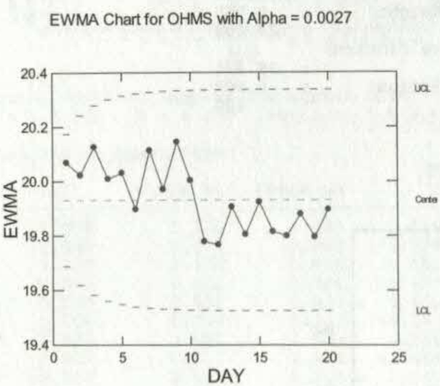
The control limits are wider for the first three samples than for the remaining samples. This is because the moving average and its control limits are based only on the first sample for day 1, on two samples for day 2, on three samples for day 3, and on four samples for each remaining day.

Example 20
EWMA Chart

The EWMA chart plots the exponentially weighted moving average of the y variable as a function of an x variable that identifies each sample. The plot has a center line and control limits. For example, to produce an EWMA chart of the *BOXES* data file, the input is:

```
QC
   USE BOXES
   EWMA OHMS*DAY / K=0.2  YMIN=19.4  YMAX=20.4
```

The output is:



There is a downward trend in the EWMA for the ohms of resistance over the days of sampling. If you examine the comparable MA chart, you see a hint of the same downward trend. The control limits on the EWMA broaden over successive samples because of the method of calculation of the standard error of the EWMA.

Example 21
X-MR Chart

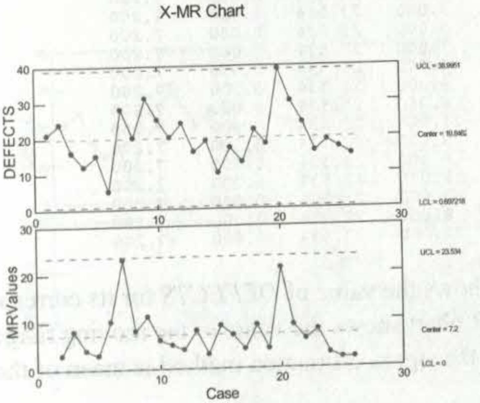
To illustrate the X-MR chart we use the *BOARDS* data. In this data, the number of defects observed in 26 successive samples of 100 printed circuit boards are listed against the corresponding sample number.

The input is:

```
QC
    USE BOARDS
    PLENGTH LONG
    XMR DEFECTS/ SLIMITS= -3 3 MEAN
```

The output is:

```
Number of Lines of Input Data Read              : 26.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 : 26.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                       : 19.846
Estimated Population Standard Deviation         :  7.165
Total N (Excluding Missing Data)                :     26
```




```
Listing of Chart Data
1    2    3    4    5    6    7    8    DEFECTS    UCL      LCL     Center
------------------------------------------------------------------------
                                       21.000   38.995   0.697   19.846
                                       24.000   38.995   0.697   19.846
                                       16.000   38.995   0.697   19.846
                                       12.000   38.995   0.697   19.846
                                       15.000   38.995   0.697   19.846
                                        5.000   38.995   0.697   19.846
                                       28.000   38.995   0.697   19.846
                                       20.000   38.995   0.697   19.846
                                       31.000   38.995   0.697   19.846
                                       25.000   38.995   0.697   19.846
                                       20.000   38.995   0.697   19.846
                                       24.000   38.995   0.697   19.846
                                       16.000   38.995   0.697   19.846
                                       19.000   38.995   0.697   19.846
                                       10.000   38.995   0.697   19.846
                                       17.000   38.995   0.697   19.846
                                       13.000   38.995   0.697   19.846
                                       22.000   38.995   0.697   19.846
                                       18.000   38.995   0.697   19.846
                                       39.000   38.995   0.697   19.846
                                       30.000   38.995   0.697   19.846
                                       24.000   38.995   0.697   19.846
```

16.000	38.995	0.697	19.846
19.000	38.995	0.697	19.846
17.000	38.995	0.697	19.846
15.000	38.995	0.697	19.846

Listing of Chart Data

1	2	3	4	5	6	7	8	MR	UCL	LCL	Center
								.	23.534	0.000	7.200
								3.000	23.534	0.000	7.200
								8.000	23.534	0.000	7.200
								4.000	23.534	0.000	7.200
								3.000	23.534	0.000	7.200
								10.000	23.534	0.000	7.200
								23.000	23.534	0.000	7.200
								8.000	23.534	0.000	7.200
								11.000	23.534	0.000	7.200
								6.000	23.534	0.000	7.200
								5.000	23.534	0.000	7.200
								4.000	23.534	0.000	7.200
								8.000	23.534	0.000	7.200
								3.000	23.534	0.000	7.200
								9.000	23.534	0.000	7.200
								7.000	23.534	0.000	7.200
								4.000	23.534	0.000	7.200
								9.000	23.534	0.000	7.200
								4.000	23.534	0.000	7.200
								21.000	23.534	0.000	7.200
								9.000	23.534	0.000	7.200
								6.000	23.534	0.000	7.200
								8.000	23.534	0.000	7.200
								3.000	23.534	0.000	7.200
								2.000	23.534	0.000	7.200
								2.000	23.534	0.000	7.200

The *X* or the individual chart shows the value of *DEFECTS* for its corresponding *SAMPLE* for 26 cases. The MR chart shows the value of the moving ranges for the corresponding *SAMPLE* when the sigma estimation method is mean of the moving range values with width 2.

From the output it is clear that only a single value (corresponding to the Sample no. 20) goes beyond the control limits as far as *DEFECTS* are concerned, and overall it is under control. But the MR Chart shows that there are a couple of very high values even though most of the MR values lie below the control lines.

Example 22
X-MR Chart (Sigma Estimation with Median)

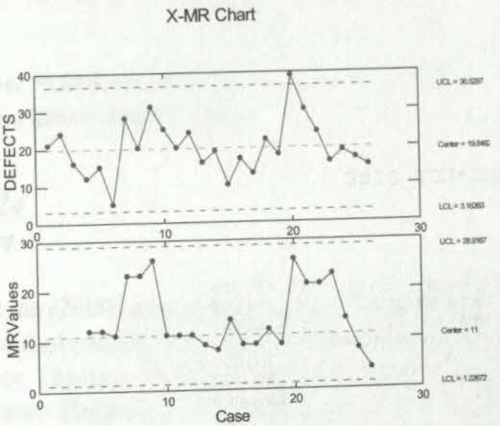
We now check the behavior of the chart using the same data file when the sigma estimation method is median of moving range values with width four.

The input is:

```
QC
   USE BOARDS
   XMR DEFECTS / WID=4    MEDIAN
```

The output is:

```
Number of Lines of Input Data Read                    : 26.000
Number with Missing Data or Zero Weight               :  0.000
Number of Samples to be Plotted                       : 26.000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                             : 19.846
Estimated Population Standard Deviation               :  7.165
Total N (Excluding Missing Data)                      :     26
```



In this output, significant changes can be figured out in the MR chart. Compared to the previous MR Chart, many of these MR values have come closer to the UCL. So with the changes in width of the subgroup size, the pattern of the MR chart changes.

Example 23
Regression Charts

The following data set is from Neter, et al. (1996). Here, a spare part is manufactured by the Westwood Company once a month. The lot size varies from month to month because of differences in demand. The data below from the *WESTWOOD* data file show the number of man-hours of labor for each of 10 lot sizes manufactured.

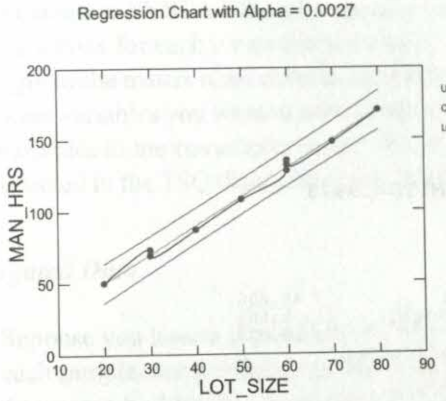
LOT_SIZE	MAN_HRS
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132

The input is:

```
QC
   USE WESTWOOD
   QCREGRESS MAN_HRS*LOT_SIZE
```

The output is:

```
Number of Lines of Input Data Read            :  10.000
Number with Missing Data or Zero Weight       :   0.000
Number of Samples to be Plotted               :  10.000
(Only Subgroups Containing Data are Plotted)
Mean of Predictor Variable (X)                :  50.000
Standard Deviation of X Variable              :  19.437
Mean of Predicted Variable (Y)                : 110.000
Standard Deviation of Y Variable              :  38.959
Estimated Regression Slope                    :   2.000
Estimated Regression Intercept                :  10.000
Standard Error of Estimate                    :   2.739
Pearson Correlation Coefficient               :   0.998
```



Use PLENGTH MEDIUM to view the predicted values (*CENTER*), as well as the upper and lower control limits.

Example 24
TSQ Chart

Ryan (2000) provides some interesting data for which a TSQ chart signals an out-of-control condition, while individual Shewhart *X* charts for the same control variables do not. The data set, *RYAN*, is partially listed below to show how to enter raw data for this chart. There are two control variables, *Y1* and *Y2*, and a sample identifier called *SAMPLE*. Each of the 20 samples consists of four cases.

SAMPLE	Y1	Y2
1	72	23
1	84	30
1	79	28
1	49	10
2	56	14
2	87	31
2	33	8
2	42	9
…	…	…
20	35	10
20	38	11
20	41	13

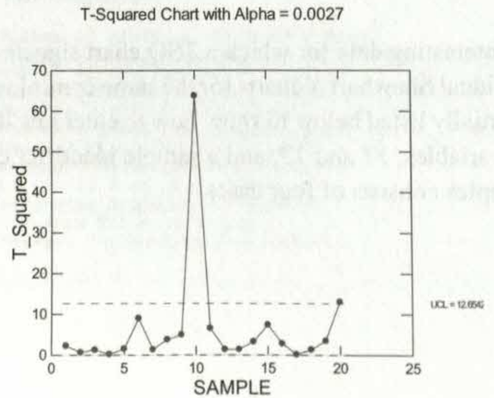
20	46	16

The input is:

```
QC
   USE RYAN
   PLENGTH SHORT
   TSQ Y1,Y2*SAMPLE/PLIMITS=.9973
```

The output is:

```
Number of Lines of Input Data Read              : 80.000
Number with Missing Data or Zero Weight         :  0.000
Number of Samples to be Plotted                 : 20.000
(Only Subgroups Containing Data are Plotted)
Total N (Excluding Missing Data)                :     80

Grand Means of Dependent Variables
Y1       Y2
60.375   18.488

Pooled Variance-Covariance Matrix

222.033
103.117   56.579
```



As you can see from the chart, samples 10 and 20 produced values of *T*-squared that were outside the upper control limit. Ryan shows that separate *X* charts for *Y1* and *Y2* do not show these samples as out of control, indicating that even though the bivariate process is out of control, the variables when considered individually does not appear to be so.

Suppose that this matrix is in a file named *MYFILE*. You would select Use from external file and click File to select the file to use. The input matrix must be a

rectangular SYSTAT file of a specific form. There must be one column (variable) in the matrix for each *y* variable selected in the TSQ dialog box. The first variable to the right of the matrix must contain the mean for each selected variable. You can have any other variables you want in subsequent columns. It does not matter what you name the variables in the covariance matrix file. However, the order in which the variables are selected in the TSQ dialog box must match the order of the variables in the matrix file.

Aggregated Data

Suppose you have a file called *RYANSAVE* that contains the means for *Y1* and *Y2* for each sample along with the sample size, *n*, for each sample. You could then select Aggregate by Sample to use these data as aggregated input. To do so, you must also create a file containing a variance-covariance matrix and grand means for *Y1* and *Y2*. To make life easy, simply copy the variance-covariance matrix from the previous output into a data file named *RYANCOV*. Here is the form in which to enter the data:

Y1	Y2	MEANS
222.033	.	60.375
103.117	56.579	18.488

When SYSTAT reads the variance-covariance matrix from the file named in the Covariance Matrix option, it ignores anything above the diagonal of the matrix. For convenience, you can designate these values as missing, as shown above. Because the matrix is symmetric, the actual value for the missing cell is 103.117 and could be entered as such, if desired. The column named *MEANS* contains the means for *Y1* and *Y2* entered in the same order as these variables are listed for the variance-covariance matrix. The variable names in the *RYANCOV* file could be anything; we named them *Y1* and *Y2* for convenience.

Now, you could analyze the aggregated data. Notice that the order of variables in the TSQ instructions matches (and must match) the order in which the data for these variables appear in the *RYANCOV* file.

The input is:

```
USE RYANSAVE
FREQ N
TSQ Y1,Y2*SAMPLE / AGG COVAR=RYANCOV, PLIMITS=.9973
```

The resulting chart is identical to the chart from the previous example. The output does not report the empirical covariance matrix because no empirical matrix can be

computed from aggregated data. That is why the COVAR (Covariance Matrix) option is required. However, the empirical means for the control variables are reported. They happen to be the same as the means in the matrix file in this example, but they do not have to be. They are not used in the calculations when the COVAR option is used, but are reported for information. The means actually used in the calculations are those in the matrix file.

Example 25
PCA with Normal Distribution

The following example uses the data file *AIAG* originating from Automotive Industry Action Group (AIAG, 1995) and taken from Breyfogle (2003). The process variable of interest in this data file is a measurement of a critical quality characteristic of 80 samples, 5 samples collected in each of 16 subgroups. Suppose that the lower specification limit (LSL) for measurement is 0.5, the upper specification limit (USL) is 0.9, and nominal or target is 0.7.

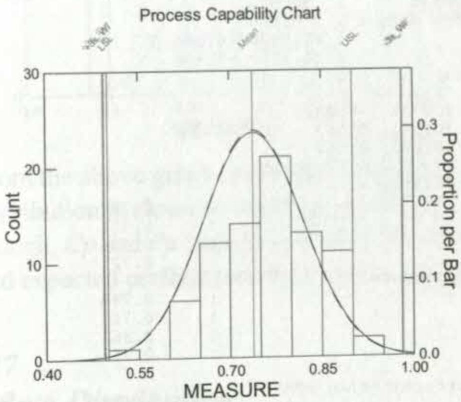
The input is

```
QC
  USE AIAG
  PLENGTH LONG
  PCA MEASURE/USL=0.9 LSL=0.5 NOMINAL=0.7 SIZE=5 DIST=NORMAL,
                 SIGMATOL=6
```

The output is:

```
Number of Lines of Input Data Read       : 80.000
Number with Missing Data or Zero Weight :  0.000
USL                                      :  0.900
Nominal                                  :  0.700
Total N (Excluding Missing Data)         :     80

Mean                          : 0.737
Standard Deviation(within)    : 0.081
Standard Deviation(overall)   : 0.082

                      Process Capability
Cp                                              :      0.822
CPU                                             :      0.668
CPL                                             :      0.976
Cpk                                             :      0.668
                      Process Performance
Pp                                              :      0.813
PPU                                             :      0.661
PPL                                             :      0.966
Ppk                                             :      0.661
Ppm                                             :      0.740
Pmk                                             :      0.661
```

```
                         Observed Performance(in PPM)
Less than LSL                                          :       0.000
More than USL                                          :       0.000
Total beyond SLs                                       :       0.000
                         Expected Capability(in PPM)
Less than LSL                                          :    1709.526
More than USL                                          :   22594.123
Total beyond SLs                                       :   24303.649
                         Expected Performance(in PPM)
Less than LSL                                          :    1882.022
More than USL                                          :   23719.778
Total beyond SLs                                       :   25601.800
```



The 3-sigma limits (sigma within and sigma overall) on the plot (labeled +3S_Wi and -3S_Wi or +3S_Ovr and -3S_Ovr) fall outside the interval defined by the specification limits (LSL and USL). That causes the index *Cp* and *Pp* to be less than 1.00 because *Cp* or *Pp* is the distance from LSL to USL divided by the distance from -3S_Wi to +3S_Wi or -3S_Ovr to +3S_Ovr. Thus the actual distribution of ohms has much of its area outside the specification limits. The *Cpk* or *Ppk* index is even smaller than *Cp* or *Pp* because it is based on the distance from the mean to +3S_Wi / +3S_Ovr (or to +3S_Wi / +3S_Ovr, which is the same distance).

Example 26
PCA With Box-Cox Transformation

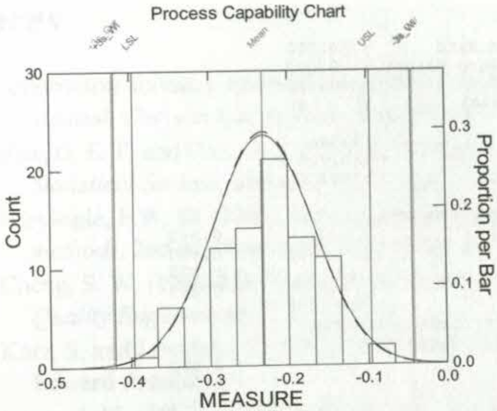
The following example uses the data file *AIAG*. From the histogram of the above example, it seems that the data distribution is skewed negatively. Here we use the Box-Cox transformation to the measurement variable hoping the *Cpk* and *Ppk* will improve.

The input is:

```
QC
  USE AIAG
  PLENGTH MEDIUM
  PCA  MEASURE / USL =0.9 LSL=0.5 SIZE=SUBGROUP DIST=NORMAL,
                 SIGMATOL=6 BOXCOX
```

The output is:

```
Optimal Value of Lambda                  :    1.800
Number of Lines of Input Data Read       :   80.000
Number with Missing Data or Zero Weight  :    0.000
USL                                      :  (0.900) -0.096
LSL                                      :  (0.500) -0.396
Total N (Excluding Missing Data)         :      80

Mean                        : (0.737) -0.232
Standard Deviation(within)  : (0.081) 0.063
Standard Deviation(overall) : (0.082) 0.064

                    Process Capability
Cp                                       :       0.797
CPU                                      :       0.721
CPL                                      :       0.874
Cpk                                      :       0.721
                    Process Performance
Pp                                       :       0.786
PPU                                      :       0.711
PPL                                      :       0.861
Ppk                                      :       0.711

                    Observed Performance(in PPM)

Less than LSL                            :       0.000
More than USL                            :       0.000
Total beyond SLs                         :       0.000
                    Expected Capability(in PPM)
Less than LSL                            :    4380.857
More than USL                            :   15296.148
Total beyond SLs                         :   19677.005
                    Expected Performance(in PPM)
Less than LSL                            :    4877.479
More than USL                            :   16504.925
Total beyond SLs                         :   21382.403
```



From the above graph of the transformed data, it is clear that the transformed data distribution is closer to a normal distribution than the original data distribution. Here though, *Cp* and *Pp* have decreased, *Cpk* and *Ppk* have increased. Expected capability and expected performance have also improved.

Example 27
PCA with Beta Distribution

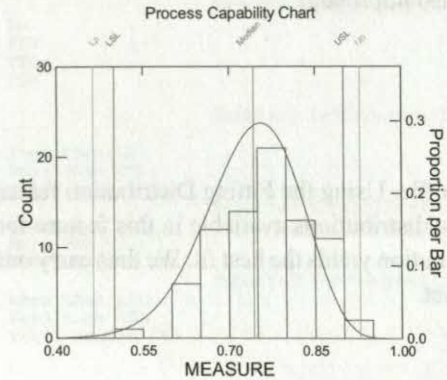
This example also uses *AIAG* data file. Using the Fitting Distribution feature of SYSTAT, we found that among the distributions available in this feature for Process Capability Analysis, the beta distribution yields the best fit. We thus carry out PCA with the beta distribution for this data set.

The input is:

```
QC
   USE AIAG
   PLENGTH MEDIUM
   PCA  MEASURE / USL =0.9 LSL=0.5 DIST=BETA SIGMATOL=6
```

The output is:

```
Number of Lines of Input Data Read       : 80.000
Number with Missing Data or Zero Weight :  0.000
USL                                      :  0.900
Total N (Excluding Missing Data)         :     80

Median                                   :0.743
Shape1(alpha)                            :20.916
Shape2(Beta)                             :7.445

                    Process Performance
                                                  : 0.862
Pp                                                : 0.850
PPU                                               : 0.870
PPL                                               : 0.850
Ppk

                    Observed Performance(in PPM)

Less than LSL                                     : 0.000
More than USL                                     : 0.000
Total beyond SLs                                  : 0.000

                    Expected Capability(in PPM)
Less than LSL                            :4348.903
More than USL                            :8972.339
Total beyond SLs                         :13321.242
```



Here *Pp* and *Ppk* have significantly increased compared to their values with the normal distribution or with the Box-Cox transformation. Total beyond specification limits (in PPM) is reduced to 13321.242 which is nearly half of the corresponding value for the normal distribution.

References

Automotive Industry Action Group (1995). *Statistical process control (SPC) reference manual*. Chrysler Corporation, Ford Motor Company, General Motors Corporation.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, 26, 211-243. (Discussion: 244-252).

Breyfogle, F.W. III (2003). *Implementing six sigma: smarter solution through statistical methods*, 2nd ed. New York: John Wiley & Sons.

Cheng, S. W. (1994-95). Practical implementation of the process capability indices. *Quality Engineering*, 7, 239-259.

Kotz, S. and Lovelace, C. (1998). *Introduction to process capability indices*. London: Edward Arnold.

Lucas, J. M. (1982). Combined Shewhart-cusum quality control schemes. *Journal of Quality Technology*, 14, 51–59.

Lucas, J. M. and Crozier, R. B. (1982). Fast initial response for cusum quality control schemes: Give your cusum a head start. *Technometrics*, 24, 199–205.

Messina,W. S. (1987). *Statistical quality control for manufacturing managers*. NewYork: John Wiley & Sons.

Montgomery, D. C. (2001). *Introduction to statistical quality control*, 4th ed. New York: John Wiley & Sons.

Nelson, L.S. (1984). The Shewhart control chart–Tests for special causes. *Journal of Quality Technology*, 16, 237-239.

Neter, J.,Wasserman,W., Kutner, M.H., and Nachtscheim, C.J. (1996). *Applied linear statistical models*. 3rd ed. New York: McGraw-Hill.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1, 239–250.

Ryan, T. P. (2000). *Statistical methods for quality improvement*, 2nd ed. New York: John Wiley & Sons.

Velleman, P. V. and Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston, MA: Duxbury Press.

Chapter

6

Random Sampling

Mangalmurti Badgujar

The random sampling procedure can be used to generate random samples from distributions that are most commonly used for statistical work. SYSTAT implements, as far as possible, the most efficient algorithms for generating samples from a given type of distribution. The procedure generates uniform random numbers, based on either the Mersenne-Twister algorithm or the Wichmann-Hill algorithm.

Mersenne-Twister (MT) is a pseudorandom number generator developed by Makoto Matsumoto and Takuji Nishimura (1998). The random seed for the algorithm can be specified by using the RSEED seed, where the seed is any integer from 1 to 4294967295 for the MT algorithm and 1 to 30000 for the Wichmann-Hill algorithm. We recommend the MT option, especially if the number of random numbers to be generated for your Monte Carlo studies exceeds 10,000.

If you would like to reproduce results involving random number generation from earlier SYSTAT versions, with an old command file or otherwise, make sure that your random number generation option (under Edit => Options => General => Random Number Generation) is Wichmann-Hill (and, of course, that your seed is the same as before).

SYSTAT's Random Sampling procedure allows the user to draw a number of samples, each of the same given size, from a distribution chosen from a list of 37 univariate distributions (discrete and continuous) with given parameters.

Statistical Background

Different methods are available for generating a random sample from a specified distribution, such as the inversion method (inversion of cumulative distribution function), the rejection method, decomposition as discrete mixtures, and the acceptance-complement method. For more information see Fishman (1996), Gentle (1998), Ross (2002), and Hörmann et al. (2004).

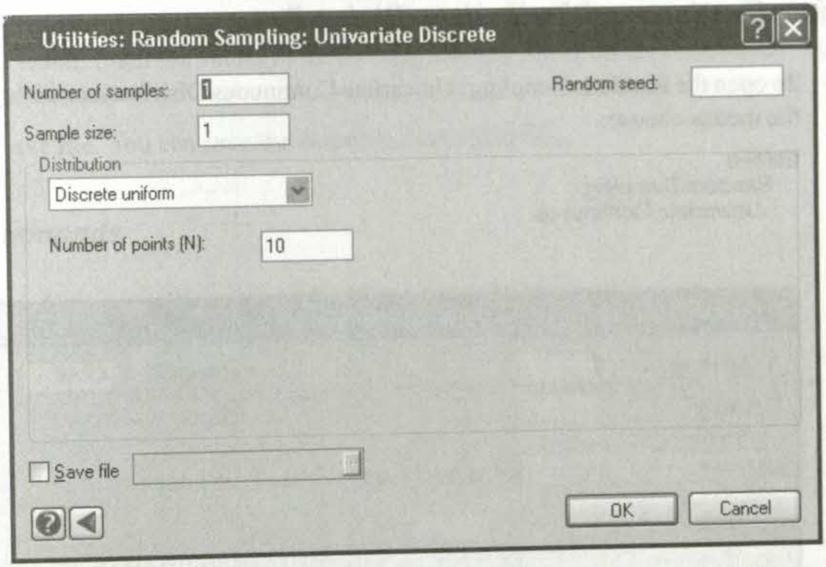
Random Sampling in SYSTAT

Before using the Random Sampling feature you should study the list of distributions, the form of the density functions, especially in respect of the parameters and the names and notations for the parameters, from the volume Data: Chapter 4: Data Transformations: Distribution Functions. It may also be useful to consult the references therein for the properties of these distributions and the meanings of the parameters. The distributions are divided into two groups -- univariate discrete and univariate continuous.

Univariate Discrete Distributions Dialog Box

To open the Random Sampling: Univariate Discrete Distributions dialog box, from the menus choose:

Utilities
  Random Sampling
    Univariate Discrete...



Number of Samples. Enter the number of samples you want to generate.

Sample size. Enter the size of the sample you want to generate.

Random seed. The default random number generator is the Mersenne-Twister (MT) algorithm. For the seed, specify any integer from 1 to 4294967295 for the MT algorithm and 1 to 30000 for the Wichmann-Hill algorithm; otherwise SYSTAT uses a seed based on system time.

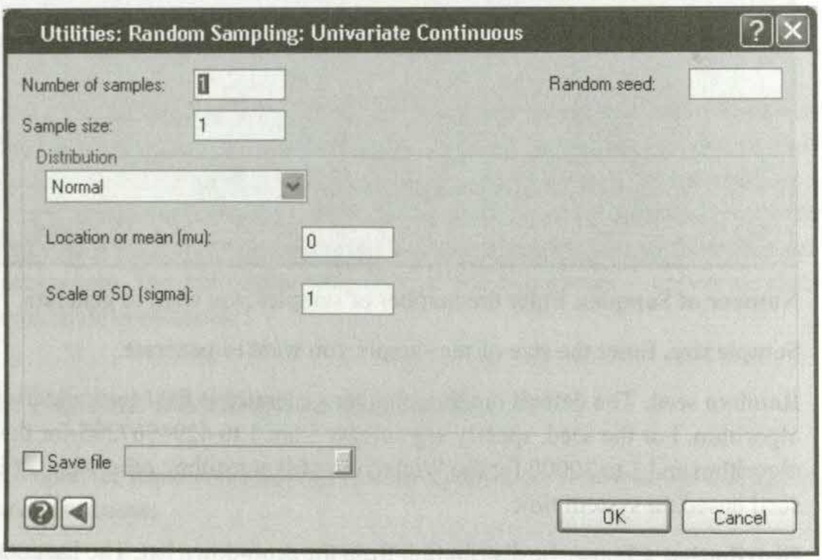
Distribution. Choose the distribution from the drop-down list. The list consists of nine univariate discrete distributions: Benford's Law, Binomial, Discrete uniform, Geometric, Hypergeometric, Logarithmic series, Negative binomial, Poisson, and Zipf. Enter the values of the parameters (depending on the distribution selected) in the box(es).

Save file. You can save the output to a specified file.

Univariate Continuous Distributions Dialog Box

To open the Random Sampling: Univariate Continuous Distributions dialog box, from the menus choose:

Utilities
  Random Sampling
    Univariate Continuous...



Number of samples. Enter the number of samples you want to generate.

Sample size. Enter the size of the sample you want to generate.

Random seed. The default random number generator is the Mersenne-Twister (MT) algorithm. For the seed, specify any integer from 1 to 4294967295 for the MT algorithm and 1 to 30000 for the Wichmann-Hill algorithm; otherwise SYSTAT uses a seed based on system time.

Distribution. Choose the distribution from the drop-down list. The list consists of twenty eight univariate continuous distributions: Beta, Cauchy, Chi-square, Double exponential (Laplace), Erlang, Exponential, F, Gamma, Generalized lambda, Gompertz, Gumbel, Inverse Gaussian (Wald), Logistic, Loglogistic, Logit normal, Lognormal, Normal, Non-central chi-square, Non-central F, Non-central t, Pareto,

Rayleigh, t, Smallest extreme value, Studentized range, Triangular, Uniform, and Weibull. Enter the values of the parameters (depending on the distribution selected) in the box(es).

Save file. You can save the output to a specified file.

Using Commands

For Univariate Discrete and Continuous random sampling:

```
RANDOMSAMP
SAVE filename
UNIVARIATE distribution notation (parameterlist)/SIZE = n1
NSAMP=n2 RSEED=n
```

The distribution notation consists of parameter values as its arguments.

Distribution Notations used in Random Sampling

Distribution Name	Distribution Notation	Parameter(s)
Benford's law	BLRN	*(B)*
Binomial	NRN	*(n,p)*
Discrete uniform	DURN	*(N)*
Geometric	GERN	*(p)*
Hypergeometric	HRN	*(N,m,n)*
Logarithmic series	LSRN	*(theta)*
Negative binomial	NBRN	*(k,p)*
Poisson	PRN	*(lambda)*
Zipf	ZIRN	*(shp)*
Beta	BRN	*(shp1,shp2)*
Cauchy	CRN	*(loc,sc)*
Chi-square	XRN	*(df)*
Double exponential (Laplace)	DERN	*(loc,sc)*
Erlang	ERRN	*(shp,sc)*
Exponential	ERN	*(loc,sc)*
F	FRN	*(df1,df2)*
Gamma	GRN	*(shp,sc)*

Distribution Name	Distribution Notation	Parameter(s)
Generalized lambda	GLRN	*(lambda1,lambda2,lambda3,lambda4)*
Gompertz	GORN	*(b,c)*
Gumbel	GURN	*(loc,sc)*
Inverse Gaussian (Wald)	IGRN	*(loc,sc)*
Logistic	LRN	*(loc,sc)*
Logit normal	ENRN	*(loc,sc)*
Loglogistic	LORN	*(logsc, shp)*
Lognormal	LNRN	*(loc,sc)*
Non-central chi-square	NXRN	*(df1,nc)*
Non-central F	NFRN	*(df1,df2,nc)*
Non-central t	NTRN	*(df.nc)*
Normal	ZRN	*(loc,sc)*
Pareto	PARN	*(sc,shp)*
Rayleigh	RRN	*(sc)*
Smallest extreme value	SERN	*(loc,sc)*
Studentized range	SRN	*(k,df)*
t	TRN	*(df)*
Triangular	TRRN	*(a,b,c)*
Weibull	WRN	*(sc,shp)*
Uniform	URN	*(a,b)*

Example: Normal random number with parameters (0,1)

```
RANDOMSAMP
UNIVARIATE ZRN (0,1)
```

Usage Considerations

Types of data. No input data are needed.

Print options. There are no print options. The generated data are shown in the data editor.

Quick Graphs. RANDSAMP produces no Quick Graphs. You use the generated file and produce the graphs you want. See the examples.

Saving files. The generated samples can be saved in the file mentioned. For all distributions the case number refers to the observation number and the column names are s1, s2, …(the number after s denotes the sample number).

BY groups. By groups is not relevant.

Case frequencies. Case frequency is not relevant.

Case weights. Case weight is not relevant.

Examples

Example 1
Sampling Distribution of Double Exponential (Laplace) Median

In this example we generate 500 samples, each of size 20, and investigate the distribution of the sample median by computing the median of each sample.

The input is:

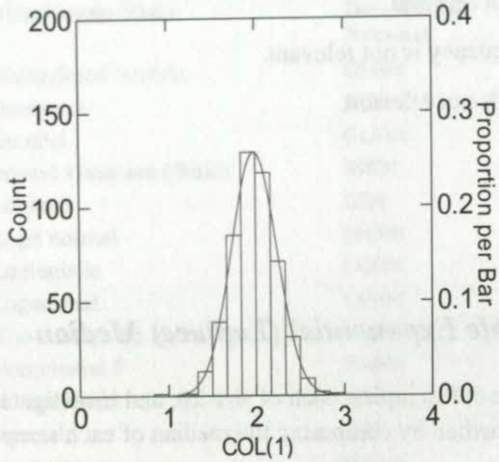
```
RANDSAMP
SAVE "DOUEXP.SYD"
UNIVARIATE DERN(2,1) / SIZE=20  NSAMP=500 RSEED=2341
```

Using the generated (500) samples, the distribution of sample median can be obtained. For this the input is:

```
USE "DOUEXP.SYD"
SSAVE 'CBSTAT.SYD'
CSTATISTICS S1..S500/MEDIAN
USE 'CBSTAT.SYD'
TRANSPOSE S1..S500
CSTATISTICS COL(1)/MAXIMUM MEAN MINIMUM SD VARIANCE N
SWtest
BEGIN
DENSITY COL(1)/HIST XMIN=0 XMAX=4
DENSITY COL(1)/NORMAL XMIN=0 XMAX=4
END
```

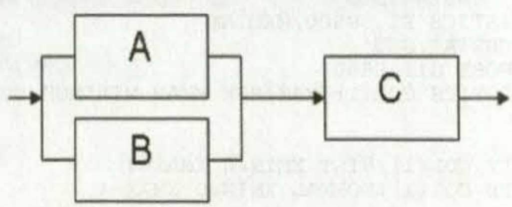
The last part of output is as follows. *COL(1)* contains the sample median:

	COL(1)
N of Cases	500.000
Minimum	0.909
Maximum	2.924
Arithmetic Mean	1.981
Standard Deviation	0.258
Variance	0.066
Shapiro-Wilk Statistic	0.996
Shapiro-Wilk p-value	0.183



We observe that the sampling distribution of the double exponential sample median can be described to be normal.

Example 2
Simulation of Assembly System



Consider a system having two parallel subsystems (A and B) connected in a series with another subsystem (C), as shown in the structural diagram. In such a system, work at "C" can start only after the work at "A" and "B" is completed. The process completion time for this system is the maximum of the processing times for "A" and "B" plus the processing time for "C". Assume that the system is a production line for a specific

product, and that the processing time distributions for the three subsystems are independent. Let us specify that:

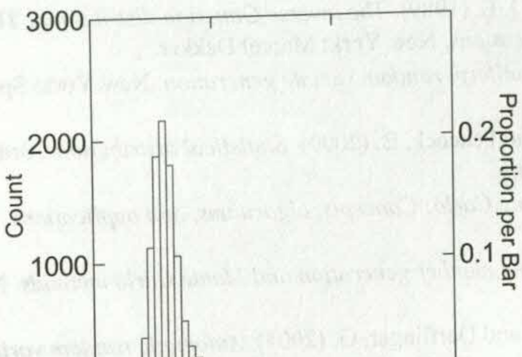
A~ Exponential (0.2, 0.7)
B~ Uniform (0.2, 1.2)
C~ Normal (2, 0.3)

The production engineer wants to find the distribution of manufacturing time and to estimate the probability that the manufacturing time is less than 5 units of time.

The input is:

```
RANDSAMP
UNIVARIATE ERN (0.2, 0.7)/SIZE = 10000 NSAMP=1 RSEED=123
LET s2=URN (0.2, 1.2)
LET s3=ZRN (2, 0.3)
LET TIME=MAX (s1, s2) +s3
DENSITY time / HIST
Let prob = (time <=5)
CSTATISTICS prob/mean
```

The output is:




```
                   | PROB
-------------------+------
Arithmetic Mean    | 0.979
```

The output shows the histogram of 10000 simulated manufacturing times; the estimated probability that manufacturing time is less than 5 time units is 0.979.

Computation

Algorithms

The algorithms used here for random sampling from specified distributions can be found in Devroye (1986), Bratley et al. (1987), Chhikara and Folks (1989), Fishman (1996), Gentle (1998), Evans et. al. (2000), Karian and Dudewicz (2000), Ross (2002), and Hörmann et al. (2004). For some distributions, the inverse CDF method (analytical or numerical); for others, certain special methods are used.

References

Bratley, P., Fox, B.L., and Schrage, L.E. (1987). *A guide to simulation*. 2nd ed, New York: Springer-Verlag.

Chhikara, R. S. and Folks, J. L. (1989). *The inverse Gaussian distribution: Theory, methodology, and applications*. New York: Marcel Dekker.

Devroye, L. (1986). *Non- uniform random variate generation*. New York: Springer-Verlag.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical distributions*. 3rd ed. New York: John Wiley & Sons.

Fishman, G.S. (1996). *Monte Carlo: Concepts, algorithms, and applications*. New York: Springer-Verlag.

Gentle, J. E. (1998). *Random number generation and Monte Carlo methods*. New York: Springer-Verlag.

Hörmann, W., Leydold, J., and Derflinger, G. (2004). *Automatic random variate generation*. Berlin: Springer-Verlag.

* Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. 3rd ed. New York: John Wiley & Sons.

* Johnson, N. L., Kotz, S., and, Balakrishnan, N. (1994). *Univariate continuous distributions*. Vol. 1, 2nd ed. New York: John Wiley & Sons.

* Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Univariate continuous distributions*. Vol. 2, 2nd ed. New York: John Wiley & Sons.

Karian, Z.A. and Dudewicz, E.J. (2000). *Fitting statistical distributions: The generalized lambda distribution and generalized bootstrap methods*. Boca Raton, FL: CRC Press.

Matsumoto, M. and Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8, 3-30.

Ross, S.M. (2002). *Simulation*, 3rd ed. San Diego, CA: Academic Press.

(* indicates additional references.)

Chapter

7

Response Surface Methods

Meghana Kulkarni and K.V.S. Rammurty

Response Surface Methods (RSM) are used to develop an empirical model, commonly called response surface, for the response of a process in terms of the relevant controllable factors. RSM determines the operating conditions that produce the optimum response. RSM in SYSTAT allows you to specify and fit a model up to second order. It also provides an option to include a block variable. SYSTAT provides regression coefficients of both coded and uncoded factors. For each response, RSM fits a model and provides the ANOVA and the 'Lack of Fit' test separately when there is more than one response. Contour and Surface plots of each response for pairs of factors are also produced.

SYSTAT offers three kinds of optimization techniques: Canonical Analysis, Ridge Analysis and Desirability Analysis. Canonical Analysis determines the stationary point and the optimal response for each response separately, and its nature (maximum or minimum or saddle point). Ridge Analysis helps you to determine the direction in which to look for the optimal response in case of saddle surface or when the stationary point is beyond the experimental region. Desirability Analysis optimizes more than one response variable simultaneously. SYSTAT provides desirability plots as Quick Graphs.

Statistical Background

Response surface methods (RSM), introduced by Box and Wilson (1951), are used in industries to explore the relationship between a response variable and several input factors, to determine the optimal settings of the factors, and to optimize the process or product. RSM are extensively used in situations where there are many input factors

that may influence one or more response variables. A typical application of RSM proceeds on the following lines:

- Identify the relevant input variables that will help to build up an appropriate response surface
- Determine the optimal settings of the factors that will yield an optimum response
- Optimize multiple responses

For the theoretical development and review of RSM, refer to Hill and Hunter (1966), Khuri and Cornell (1996) and Myers and Montgomery (2002). The following sections briefly discuss how RSM in SYSTAT carries out the above steps.

Fitting a Response Surface

Suppose in a certain chemical industry, different chemical reactions are performed. These reactions are often influenced by the process temperature, pressure, and, sometimes, time of reaction. You may be interested in finding the combination(s) of the levels of temperature, pressure, and time that give(s) the best yield from the reaction. To do so, you need to explore the relation between the factors and the response variable. Let y denote the yield for the reaction.

A response surface model is

$$y = g(x_1, x_2, \ldots, x_k) + \varepsilon$$

where $x_1, x_2, \ldots, x_k$ are k controllable factors (input variables), and ε is the noise factor. The surface represented by the above function is called a *response surface*. Generally the form of g is not known, and is perhaps too complicated to fit. So we approximate it with a low order polynomial in the neighborhood of the optimum response. Quite often a second-order model is considered to be accurate enough while approximating a surface in a small region. The full second order model is

$$y = \beta_0 + x'\beta + x'Bx + \varepsilon = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \sum_{j=1}^{k} \beta_{ij} x_i x_j + \varepsilon$$

If you have information that some of the terms of the above model are not important, you can remove such terms from the model. In SYSTAT, you can fit a variety of models : Full Model, Linear, Linear + Interaction, Linear + Quadratic.

For interpreting the results of the optimization techniques, the values of different input variables should be comparable. If the input variables are expressed in natural units like degrees Celsius (^0C), grams per liter etc., it is difficult to compare them. So, it is always advisable to code them to unit free variables $x_1, x_2, \ldots, x_k$. By default SYSTAT codes each input variable within [-1 1] by defining coded value 'C' as

$C = 2 *(O-M) / D$

where O is the original value, M is the average of the upper value and lower value of the variable and D is their range. SYSTAT computes the estimates of regression coefficients for both coded and uncoded data.

Testing Lack of Fit

To assess the adequacy of the fitted respone surface model, SYSTAT performs lack of fit test. Lack of fit test is performed when data contain repeated observations. The Sum of Squares due to error (SS_E) is divided into two parts: one is Sum of Squares due to lack of fit (SS_{LOF}), the other is Sum of Squares due to pure error (SS_{PE}). So we have:

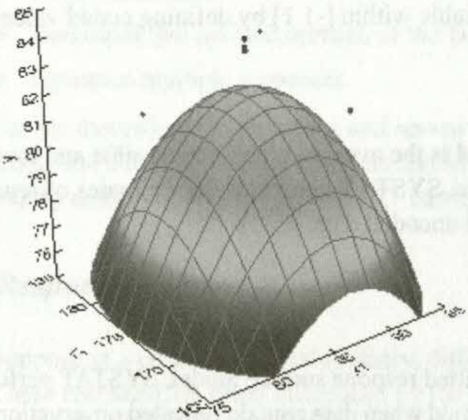
$$SS_E = SS_{PE} + SS_{LOF}$$

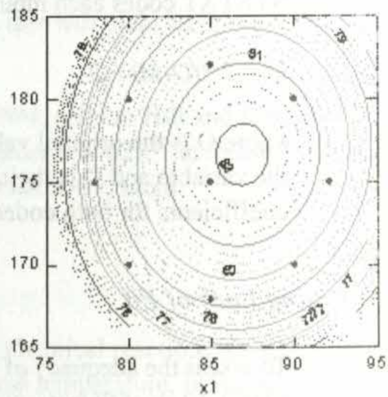
The F-ratio for lack of fit is computed by comparing SS_{PE} and SS_{LOF}. If the ratio is significant, it tells us that the fitted model is inadequate and appropriate measures should be taken to find out where the inadequacy occurs and how to avoid it.

Contour and Surface plot

In SYSTAT, you can view the fitted response graphically, either in the three-dimensional space (surface plot) or as contour plot; these help to visualize the shape of the response surface. Contour plot is a two-dimensional graph that shows contours of constant response with the axis system being specific pair of factors, say x_i and x_j. Contour plot is very useful to visualize the location of the stationary point and the nature of the response system. When the model contains more than two factors, you cannot have a contour plot or a surface plot of all the variables at once. In such a situation, you can plot a pair of factors at a time holding the remaining factors constant.

The following are the surface plot (left) and the contour plot (right) for a typical two-factor second order model.





Response Optimization

After establishing the relationship between response (yield in above example) and factors (process temperature, pressure, and reaction time), you can determine the factor setting that produces the optimum response. To optimize the process, SYSTAT provides three optimization techniques : *Canonical Analysis, Ridge Analysis*, and *Desirability Analysis*.

Canonical Analysis

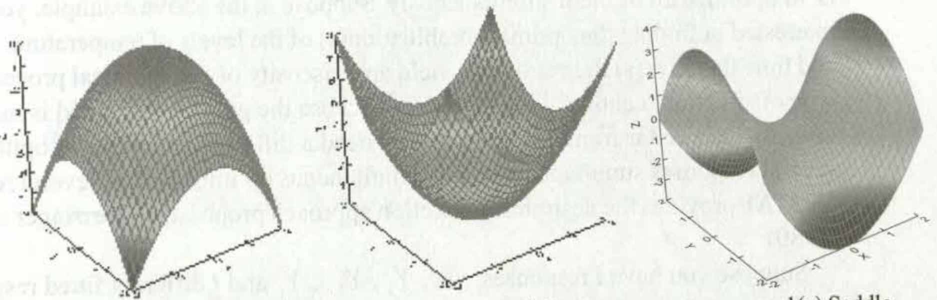
Suppose your fitted model is

$$\hat{y} = b_0 + \mathrm{x}'b + \mathrm{x}'\hat{B}\mathrm{x}$$

where, b_0, b and $\hat{B}$ are the estimated regression coefficients. SYSTAT offers Canonical Analysis to determine the factor setting that gives the optimum response. In Canonical Analysis, you can find the setting by differentiating the fitted model with respect to independent factors and equating it to the null vector. This gives you the

stationary point, i.e.,optimal factor setting. By using stationary point in the above fitted model, you can determine the optimum response (Myers and Montgomery, 2002).

At the stationary point, the response may be maximum or minimum. Sometimes it is neither maximum nor minimum but a saddle point. The geometric nature of the second-order response is displayed in Figures 1(a), 1(b) and 1(c). In Figure 1(a) the stationary point is a point of maximum response, in Figure 1(b) the stationary point is a point of minimum response and in Figure 1(c) it is a saddle point.



1(a) Maximum 1(b) Minimum 1(c) Saddle

The nature of optimality at the stationary point can be determined from the signs of the eigenvalues of the matrix $\hat{B}$. If all of the eigenvalues are negative, the stationary point is a point of maximum response. If all of the eigenvalues are positive, the stationary point is a point of minimum response. If some of the eigenvalues are positive and some are negative, the stationary point is a saddle point.

Ridge Analysis

If the optimum response obtained beyond the experimental region or the stationary point is a saddle point, you can use SYSTAT's Ridge Analysis to find the direction in which to look for the optimum response or saddle point.. In Ridge Analysis, you can find a locus of points, each of which is a point of the optimum response, with a constraint that the point is on a sphere of a certain radius. This approach also helps to get an idea about the change in response with respect to change in factor values (Myers and Montgomery, 2002).

The starting point X_0 of the radius has coordinates equal to the midpoint of values of the factors in the design (by default). Since in SYSTAT we code the data within -1 and +1, our default start point is (0, 0,…,0). The default increment of the radius is

taken as 0.1, i.e., the radii are 0, 0.1… 0.9, 1. SYSTAT provides the optimized response in the ridge table along with the un-coded values of the factors at which the response gets optimized at each radius, together with confidence intervals of the estimated response.

Desirability Analysis for optimizing more than one response

In most practical applications, data are available on several responses and the objective is to optimize all of them simultaneously. Suppose in the above example, you may be interested in finding the optimal combination(s) of the levels of temperature, pressure, and time that give(s) the maximum yield and viscosity of the chemical process. In this type of situation, canonical analysis fails because the point where yield is maximum, viscosity can be far from optimum. So we need a different technique to optimize several responses simultaneously. For simultaneous optimization of several responses, SYSTAT provides the desirability function approach proposed by Derringer and Suich (1980).

Suppose you have *l* responses, viz., Y_1 , $Y_2 \ldots Y_l$ and *l* different fitted response surfaces one for each of these responses. In the desirability function approach, you need to transform these responses into desirability functions. If you want to maximize the response Y_i , the individual desirability function d_i can be defined as:

$$d_i = \begin{cases} 0 & Y_i < L \\ \left(\dfrac{Y_i - L}{T - L}\right)^r & L \le Y_i \le T \\ 1 & Y_i > T \end{cases}$$

The value of d_i lies between 0 and 1.Practical experience shows that any value of Y_i below some lower value *L* is not acceptable, and that an ideal target value of Y_i would be *T*. In the above equation, *r* is some weight specified depending upon how rapidly you want the desirability to move towards 1. If the response is well accepted only if it is close to *T*, then specify *r* as a value greater than 1. In case any response just above *L* is well accepted, then specify *r* as a value less than 1.

Similar arguments can show that if you want to minimize Y_i and keep it always above some upper value *U*, then the desirability function becomes

$$d_i = \begin{cases} 1 & Y_i < T \\ \left(\dfrac{U - Y_i}{U - T}\right)^r & T \le Y_i \le U \\ 0 & Y_i > U \end{cases}$$

If the target for Y_i is to keep between lower (L) and Upper (U) limits, then

$$d_i = \begin{cases} 0 & Y_i < L \\ \left(\dfrac{Y_i - L}{T - L}\right)^r & L \le Y_i \le T \\ \left(\dfrac{U - Y_i}{U - T}\right)^r & T \le Y_i \le U \\ 0 & Y_i > U \end{cases}$$

The next step is to combine all these d_i's into a single index.This is done by defining the overall desirability D (the weighted geometric mean of all d_i's with 'importance' s_i) is

$$D = \left(\prod_{i=1}^{l} d_i^{\,s_i} \right)^{1 \Big/ \sum_{i=1}^{l} s_i}$$

SYSTAT provides the overall desirability D and desirability plots as Quick Graphs.

Response Surface Methods in SYSTAT

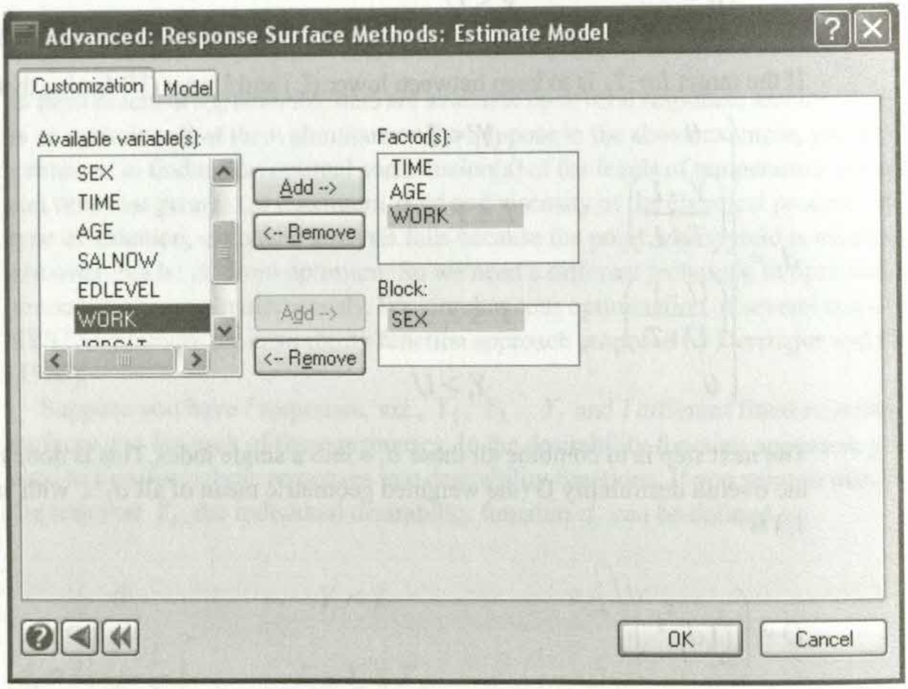
Response Surface Methods: Estimated Model Dialog Box

To open the Response Surface Methods: Estimate Model dialog box, from the menus choose:

Advanced
  Response Surface Methods
    Estimate Model...

Customization

To specify the factors and a block variable, click the Customization tab in the Response Surface Methods: Estimate Model dialog box.

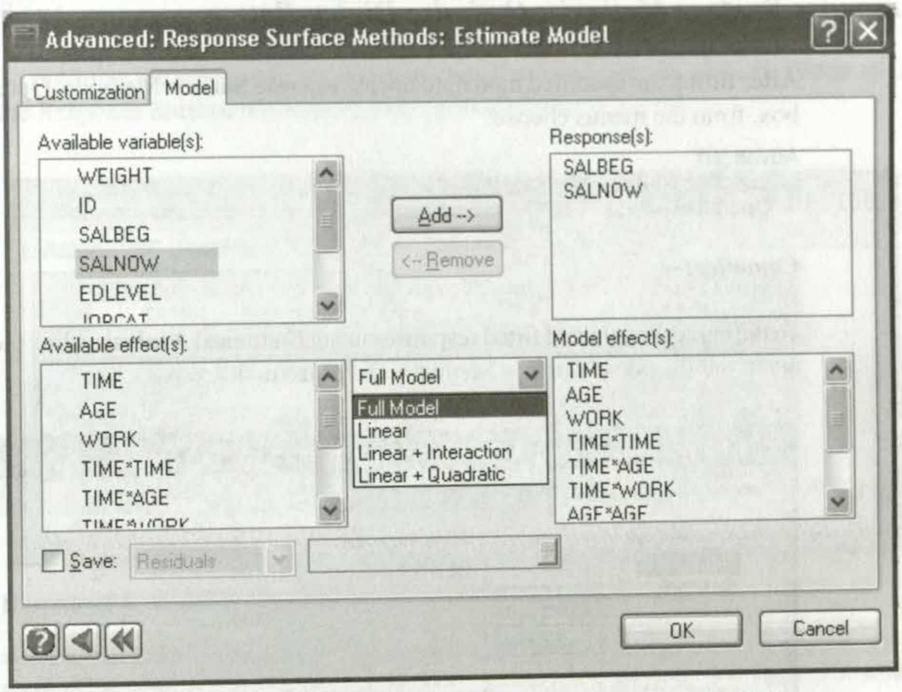


Factor(s). Select one or more predictor variables. The factors should be numeric.

Block. Select a blocking variable.

Model

After specifying the factors in the Customization ***tab, to specify the response variables and a response surface model, click the*** Model ***tab in the Response Surface Methods: Estimate Model dialog box.***



Response(s). Select the response variable(s) you want to fit the model for.

Model effect(s). Build the model by adding main effects and interaction terms from the Available effect(s) list. For instance, if A and B are selected as factors in Customization tab then A, B, A*A, B*B, and A*B are available. Alternatively, you can choose the model from the drop-down list of model types. The default is Full Model. However, the choice of effects is restricted by the condition that if a quadratic term or an interaction term is included in the model then the corresponding linear terms are also to be included. For instance, if B*B is chosen then B has to be chosen ; if B*C is chosen then B and C have to be chosen.

Save. You can save the following results to a file:

- **Residuals.** Saves the residuals and the predicted values.
- **Residuals/Data.** Saves the residuals along with data.
- **Coefficients.** Saves the estimates of the regression coefficients.

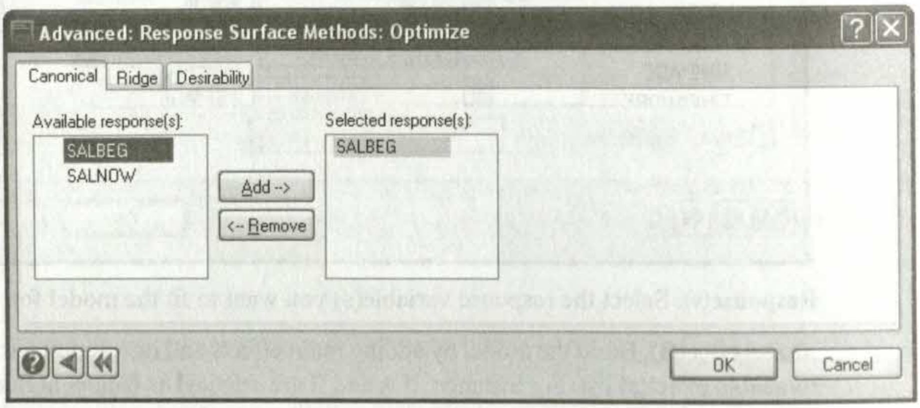
Response Surface Methods: Optimize Dialog Box

After fitting the specified model, to open Response Surface Methods: Optimize dialog box, from the menus choose:

Advanced
 Response Surface Methods
  Optimize...

Canonical

To optimize the selected fitted responses using Canonical Analysis, click the Canonical tab in the Response Surface Methods: Optimize dialog box.

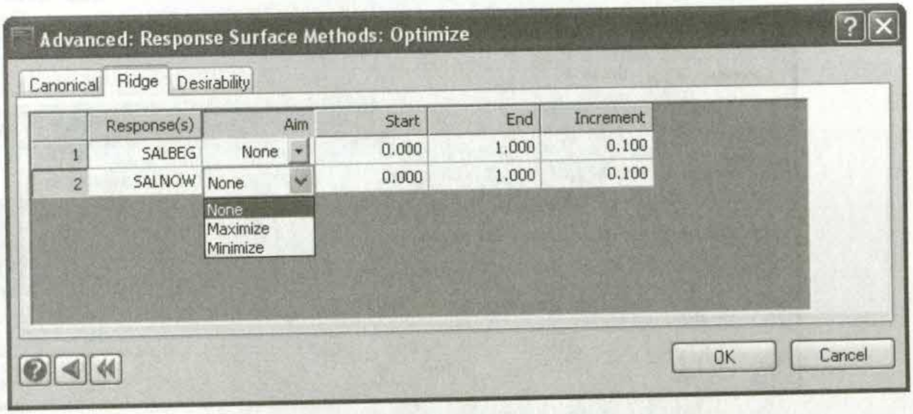


Available Response(s). Displays the response(s) selected in the Model tab.

Selected response(s). Select the response(s) you want to optimize.

Ridge

To optimize the selected fitted responses using Ridge Analysis, click the Ridge tab in the Response Surface Methods: Optimize dialog box.



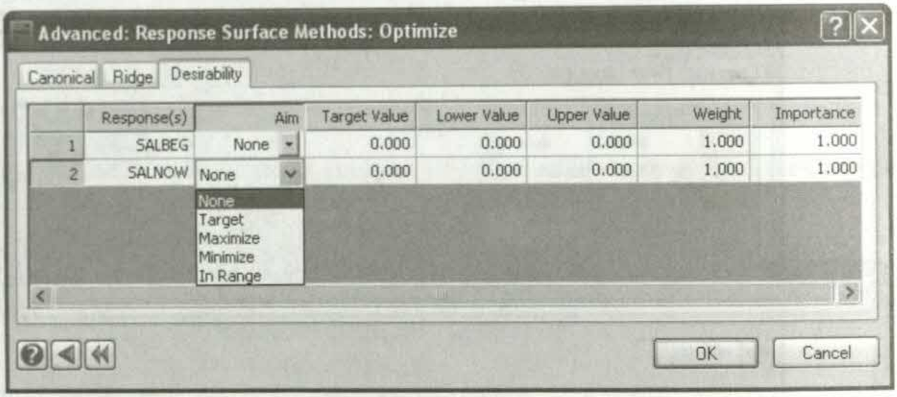
Response(s). Displays the response(s) selected in the Model tab.

Aim. Specify whether you want to Maximize or Minimize a response. If you do not want ridge analysis for any fitted response, specify aim as None. The default is None.

- **Start.** Specify starting value of the radius at which the ridge is computed.The value should be in [0,5). The default is 0.
- **End.** Specify the end value of the radius at which the ridge is computed. The value should be greater than the starting point and should be less than or equal to 5. The default is 1.
- **Increment.** Specify the increment between two consecutive radii. The default is 0.1.

Desirability

To optimize the fitted responses simultaneously, click the Desirability tab in the Response Surface Methods: Optimize dialog box.



Aim. For each response, choose one of the options available in the Aim. Your aim can be Maximize, Minimize, Target, In Range, or None :

- **Maximize.** Specify Lower Value and Target Value.
- **Minimize.** Specify Upper Value and Target Value.
- **Target.** Specify Lower Value, Target Value and Upper Value.
- **In Range.** Specify Lower Value and Upper Value. If Aim is In Range, no Desirability Analysis will be calculated for that response.

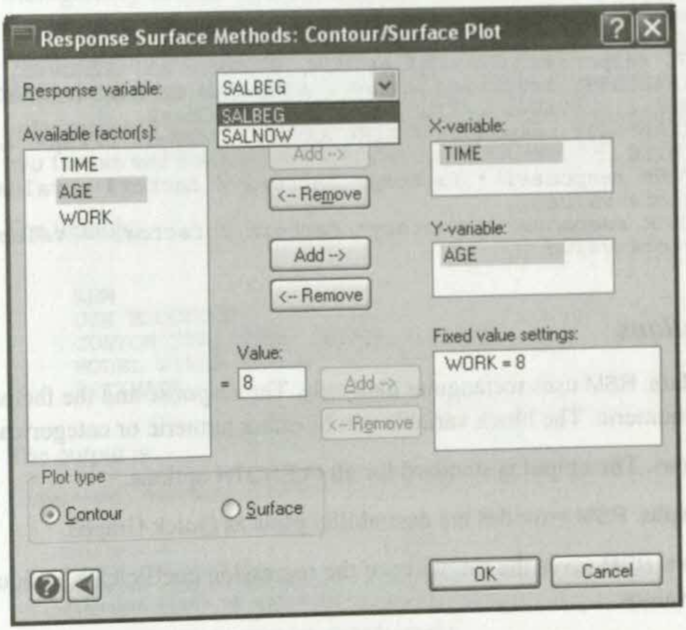
Weight. Specify the weight for each response. It can be any real number in (0,10]. The default is 1.

Importance. Specify the importance for each response.It can be any real number in (0,10]. The default is 1.

Response Surface Methods: Contour/Surface Plot Dialog Box

To open the Response Surface Methods: Contour/Surface Plot dialog box, from the menus choose:

Advanced
  Response Surface Methods
    Contour/Surface Plot...



Response variable. Choose the response variable for which you want the contour/surface plot(s).

Available factor(s). Displays the Factor(s) selected in the Customization tab.

X-variable. Select a variable to be plotted on x-axis.

Y-variable. Select a variable to be plotted on y-axis.

Fixed value settings. Fix the values of the factor(s) other than the variables to be plotted. Select the factor(s) from the Available factor(s) and enter their values in Value.

Plot type. Select the type of plot you want. The default is Contour plot.

Using Commands

Select the data with USE filename and continue with:

```
RSM
CUSTOM factor1, factor2,…/ BLOCK= blockvar
MODEL responsevarlist = factor1 + factor2 + factor1*factor2 +
  factor1*factor1+...
SAVE file name/ COEF RESID DATA
ESTIMATE

CANONICAL <no argument> or responsevarlist
RIDGE response1/AIM =MAX or MIN, RSTART= s1, REND=s2, RSTEP=s3
DESIRABILITY response1 = tar / AIM= MAX or MIN or TARGET,
  LOWER=l, UPPER=u, WEIGHT=w1, IMPORTANCE=w2
DESIRABILITY response2 / AIM= RANGE, LOWER=l, UPPER= u, WEIGHT=w
OPTIMIZE
CONTOUR response1 * factory* factorx / factor3 = value,
  factor4=value...
SURFACE response2 * factory* factorx / factor3 = value,
  factor4=value...
```

Usage Considerations

Types of data. RSM uses rectangular data only. The response and the factor variables should be numeric. The block variable can be either numeric or categorical.

Print options. The output is standard for all PLENGTH options.

Quick Graphs. RSM provides the desirability plots as Quick Graphs.

Saving files. RSM saves the estimates of the regression coefficients, residuals and predicted values.

BY groups. BY groups analysis is not available in RSM.

Case frequencies. RSM uses the FREQUENCY variable, if present, to duplicate cases.

Case weights. RSM uses the values of any WEIGHT variable to weight each case.

Examples

Example 1
Fitting a Second Order Response Surface

In this example, we use the BLOCKCCD data file, which is taken from Myers and Montgomery (2002). In a chemical process, an analysis was performed with two factors, viz., time (*TIME*) and temperature (*TEMP*), using the central composite design where two different batches of raw materials (*BLOCK*) were used. The experimenter divided the units into two blocks according to the raw materials used. We fit a second order response surface to response variable (*YIELD*) of the chemical process with the two factors and batch of raw material as a block variable.

The input is:

```
RSM
USE BLOCKCCD
CUSTOM TIME TEMP /BLOCK = BLOCK
MODEL YIELD
ESTIMATE
```

The output is:

```
Dependent Variable : YIELD
N                  :      14

  Multiple R                  : 0.999
  Squared Multiple R          : 0.998
  Adjusted Squared Multiple R : 0.996
  Standard Error of Estimate  : 0.163
```

Estimates of the Regression Coefficients

Effect	Block	Coefficient	Standard Error	t	p-value
CONSTANT		81.867	0.067	1228.863	0.000
TIME		1.319	0.082	16.162	0.000
TEMP		0.817	0.082	10.013	0.000
TIME*TIME		-2.616	0.120	-21.786	0.000
TEMP*TEMP		-1.866	0.120	-15.541	0.000
TIME*TEMP		0.250	0.163	1.532	0.169
BLOCK	1	2.229	0.044	51.103	0.000
	2	-2.229	0.000	.	.

SYSTAT computes the estimates of regression coefficients for both coded and uncoded data. It codes each factor in [-1 1].

Confidence Interval of the Regression Coefficients

Effect	Block	Coefficient	95.00% Confidence Interval Upper	Lower
CONSTANT		81.867	81.709	82.024
TIME		1.319	1.126	1.512
TEMP		0.817	0.624	1.010
TIME*TIME		-2.616	-2.900	-2.332
TEMP*TEMP		-1.866	-2.150	-1.582
TIME*TEMP		0.250	-0.136	0.636
BLOCK	1	2.229	2.126	2.332
	2	-2.229		

Analysis of Variance

Source	df	Type I SS	Mean Squares	F-ratio	p-value
Block	1	69.531	69.531	2611.095	0.000
Regression	5	27.479	5.496	206.385	0.000
Linear	2	9.626	4.813	180.734	0.000
Quadratic	2	17.791	8.896	334.054	0.000
Interaction	1	0.062	0.062	2.347	0.169
Residual Error	7	0.186	0.027		
Total Error	13	97.197			

Lack of Fit Test

Source	df	SS	Mean Squares	F-ratio	p-value
Lack of Fit	3	0.053	0.018	0.531	0.685
Pure Error	4	0.133	0.033		
Residual Error	7	0.186	0.027		

Lack of fit test indicates the adequacy of the fitted response surface model. The regression coefficients for uncoded data are

Regression Coefficients for Uncoded Factors

Effect	Block	Coefficient
CONSTANT		-1401.471
TIME		8.210
TEMP		12.759
TIME*TIME		-0.052
TEMP*TEMP		-0.037
TIME*TEMP		0.005
BLOCK	1	2.229
	2	-2.229

Note that the t statistic and p-values will be same for coded and uncoded data.

Example 2
Optimizing Response using Canonical Analysis

To determine the optimal setting of *TIME* and *TEMP* that gives optimum *YIELD*, we use Canonical Analysis.

The input is:

```
RSM
USE BLOCKCCD
CUSTOM TIME TEMP /BLOCK = BLOCK
MODEL YIELD
ESTIMATE
CANONICAL YIELD
```

The output is:

```
Canonical Analysis

       ¦  Stationary Point
Factor ¦   Coded    Uncoded
-------+-------------------
TIME   ¦   0.263     86.861
TEMP   ¦   0.236    176.672

Optimal response = 82.137 with 95% confidence interval (81.980,82.293)

Eigenvalues and Eigenvectors

               Eigenvectors
Eigenvalues     TIME     TEMP
-----------------------------
     -2.637    0.987   -0.160
     -1.846    0.160    0.987

Stationary point is Maximum.
```

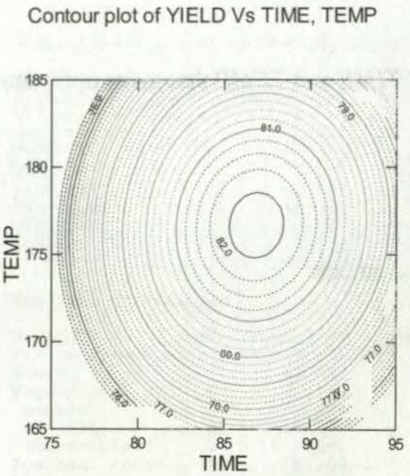
Thus, you will get the maximum *YIELD* when the settings are *TEMP* = 176.672 and *TIME* = 86.861.

We can view the fitted response surface graphically, using Contour and Surface plots.

The input is:

```
CONTOUR YIELD * TEMP * TIME
```

The output is:

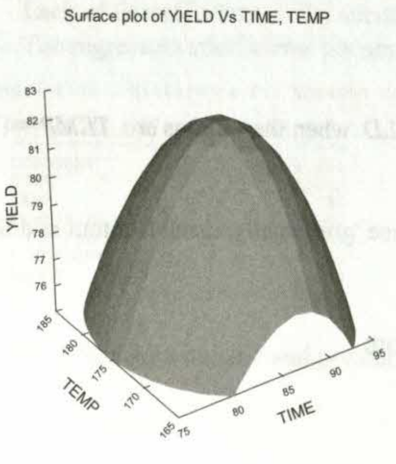


Contour plot of *YIELD* vs *TEMP* and *TIME* shows contours of constant response.

The input is:

```
SURFACE YIELD * TEMP * TIME
```

The output is:



From the surface plot, you can see that the *YIELD* is maximum around 82.

Example 3
Ridge Analysis

A step in the production of a particular polyamide resin is the addition of amines. It was felt that the manner of addition has a profound effect on the molecular weight distribution of the resin. Three variables are known to play a major role: temperature at the time of addition (^0C) (*TIME*), agitation (rpm) (*AGITATION*), and the rate of addition (min^{-1}) (*RATE*). Three levels were chosen for each factor and a Box-Behnken design was used for this experiment. The viscosity of the resin (*VISCOSITY*) was recorded as an indirect measure of the molecular weight. These data are stored in the BBD file (Myers and Montgomery, 2002). A second order response surface model is fitted to the data; it is observed that the stationary point is a saddle point.

The input is:

```
RSM
USE BBD
CUSTOM TEMP AGITATION RATE
MODEL VISCOSITY
ESTIMATE
CANONICAL VISCOSITY
```

The output is:

```
Canonical Analysis

          ¦  Stationary Point
Factor    ¦    Coded    Uncoded
----------+--------------------
TEMP      ¦    1.593    214.820
AGITATION ¦   -0.742      5.645
RATE      ¦    1.910     29.550

Optimal response = 59.555 with 95% confidence interval (36.201,82.908)

Eigenvalues and Eigenvectors

                         Eigenvectors
Eigenvalues       TEMP    AGITATION    RATE
-------------------------------------------
      2.016     -0.036      0.992     0.119
      0.301      0.587     -0.075     0.806
    -11.442     -0.809     -0.099     0.580

Stationary point is a Saddle Point.
```

The above Canonical Analysis indicates that the response surface is a saddle surface. Since the estimated surface does not have unique optimum, we use Ridge Analysis that gives the direction in which to look for maximum *VISCOSITY*.

The input is:

```
RIDGE VISCOSITY /AIM = MAX RSTART = 0 REND = 1 RSTEP = 0.1
```

The output is:

Ridge Analysis for Maximizing VISCOSITY

	Estimated	95.00% Confidence Interval		Uncoded Factor Values		
Coded Radius	Response	Upper	Lower	TEMP	AGITATION	RATE
0.000	62.000	57.248	66.752	175.000	7.500	20.000
0.100	62.345	57.608	67.081	175.127	7.711	19.734
0.200	62.687	57.992	67.381	174.754	7.949	19.564
0.300	63.053	58.423	67.683	174.303	8.194	19.450
0.400	63.451	58.896	68.006	173.870	8.443	19.372
0.500	63.883	59.399	68.368	173.473	8.693	19.320
0.600	64.352	59.911	68.793	173.114	8.945	19.286
0.700	64.858	60.404	69.312	172.788	9.197	19.268
0.800	65.402	60.850	69.955	172.491	9.449	19.260
0.900	65.985	61.217	70.753	172.219	9.702	19.262
1.000	66.607	61.486	71.727	171.967	9.955	19.271

Example 4
Multiple Response Optimization using Desirability Analysis

In the dataset MULTIRESP, we have the record of yield (*YIELD*), viscosity (*VISCOSITY*), and the average molecular weight (*MOLWEIGHT*) of a chemical process. We want to find the conditions on time (*TIME*) and temperature (*TEMP*) so that the yield and the viscosity are maximum and molecular weight will be in the range (2900, 4000) simultaneously. The canonical analysis does not provide simultaneous optimization of the three responses. We opt for the desirability analysis.

The input is:

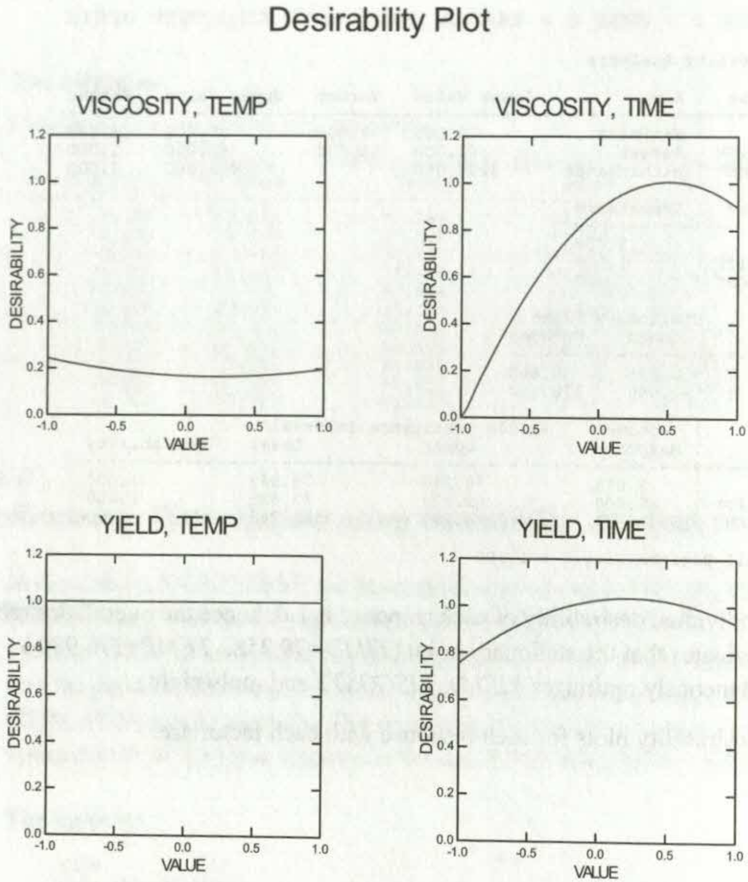
```
RSM
USE MULTIRESP
CUSTOM TIME TEMP
MODEL YIELD VISCOSITY MOLWEIGHT
ESTIMATE
DESIR YIELD = 77 /AIM = TARGET LOWER = 70 UPPER = 80 WEIGHT=1
IMPORTANCE=1
DESIR VISCOSITY =65 /AIM=MAX LOWER = 55
DESIR MOLWEIGHT /AIM =RANGE LOWER = 3200 UPPER = 3400
OPTIMIZE
```

The output is:

```
Desirability Analysis

Response  | Aim              Lower Value   Target   Upper Value   Weight
----------+-------------------------------------------------------------
YIELD     | Maximize              70.000   77.000        80.000    1.000
VISCOSITY | Target                62.000   65.000        68.000    1.000
MOLWEIGHT | WithinRange         3200.000        .      3400.000    1.000

Response  | Importance
----------+-----------
YIELD     |      1.000
VISCOSITY |      1.000
MOLWEIGHT |      1.000

       | Stationary Point
Factor |   Coded   Uncoded
-------+------------------
TIME   |   0.235    86.663
TEMP   |  -0.680   170.192

          |  Optimal     95.00% Confidence Interval
          | Response           Upper           Lower   Desirability
----------+---------------------------------------------------------
YIELD     |   78.618          78.286          78.949          1.000
VISCOSITY |   65.000          62.172          67.828          1.000
MOLWEIGHT | 3348.537        3134.289        3562.785              .

Over all Desirability D = 1.000
```

The individual desirability of each response is 1.0; hence the overall desirability being 1.0 indicates that the stationary point (*TIME* =79.358, *TEMP*=176.980) simultaneously optimizes *YIELD*, *VISCOSTY* and molweight .

The desirability plots for each response with each factor are:



Computation

To maximize the overall desirability function, Powell's algorithm is used. For details see Krishnamurthy and Sen (2001).

References

Box, G.E.P. and Wilson, K.B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series* B, 13, 1-45.

Derringer, G. and Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology*, 12, 214-219.

Hill, W.J. and Hunter, W.G. (1966). A review of response surface methodology: A literature review. *Technometrics*, 8, 571-590.

Khuri, A.I. and Cornell, J.A. (1996). *Response surfaces*. 2nd ed. New York: Marcel Dekker.

Krishnamurthy, E. V. and Sen, S. K. (2001). *Numerical algorithms*. New Delhi: EWP.

Myers, R.H. and Montgomery, D.C. (2002). *Response surface methodology*. 2nd ed. New York: John Wiley & Sons.

*Montgomery, D. C. (2000). *Design and analysis of experiments*. 5th ed. New York: John Wiley & Sons.

*Myers, R.H. and Montgomery, D.C. (2004). Response surface methodology: A retrospective and literature. *Journal of Quality Technology*, No 1, Vol. 36, Pg. 62-65.

*Ronald Christensen (2001). *Advanced linear modeling*. 2nd ed. New York: Springer-Verlag.

(* indicates additional references)

Chapter

8

Robust Regression

Shubhabrata Das, Soumyajit Ghosh, Ravindra Jore, and S.R.Kulkarni
(Revised by Sayyad Nisar Badashah and Meghana Kulkarni)

The Robust Regression (ROBREG) feature provides the following most commonly used procedures for fitting a multiple linear regression model when your data set contains outliers:

- Least Absolute Deviation (LAD) regression
- M regression
- Least Median of Squares (LMS) regression
- Least Trimmed Squares (LTS) regression
- Scale (S) regression
- Rank regression

In addition, you can use SYSTAT's Robust option in Nonlinear Models to perform LAD regression and M regression.

LAD and M regression procedures can be used when the outliers are present in the response direction while LMS, LTS, S, and Rank regression can be used when the outliers are present in both the response and the prediction directions.

In SYSTAT, LAD regression uses two methods for estimation: Iteratively Reweighted Least-Squares (IRLS) and a modified version of the simplex algorithm (Birkes and Dodge, 1993).

For M regression, you can use nine different weight functions to downweight the influence of outliers: Huber, trim, Hampel, t, Bisquare, Ramsay, Andrews, Tukey, and the n^{th} power of the absolute value of the residuals. LMS regression provides two

search procedures: Quick Search and Exhaustive Search. For large data set Quick Search method is preferred.

LTS regression implements a computationally efficient algorithm called FAST-LTS developed by Rousseeuw and Van Driessen (2000). For S regression, the SURREAL algorithm (SUfficiently Reliable REgression ALgorithm) proposed by Ruppert (1992) is implemented. It uses Tukey's biweighed function to weight the cases.

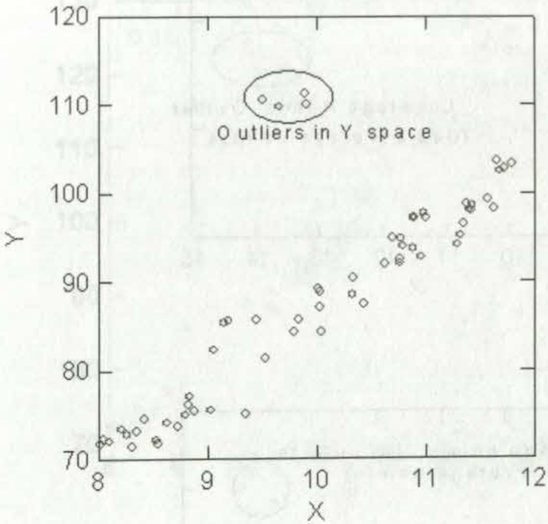
ROBREG reports robust R^2 and scale estimate of residuals. ROBREG weights cases based on residuals. For LAD (with IRLS) and M regression, weights are continuous while for the remaining procedures the weights are 0 (outlier) or 1 (not an outlier). Except for LAD (with IRLS) and M regression, SYSTAT performs the ordinary least-squares regression on outlier-free data. ROBREG allows you to save residuals/data, coefficients, predicted values, and weights. Except for LAD (with IRLS) and M regression, coefficients and residuals along with the predicted values from ordinary least-squares regression are also saved.

Statistical Background

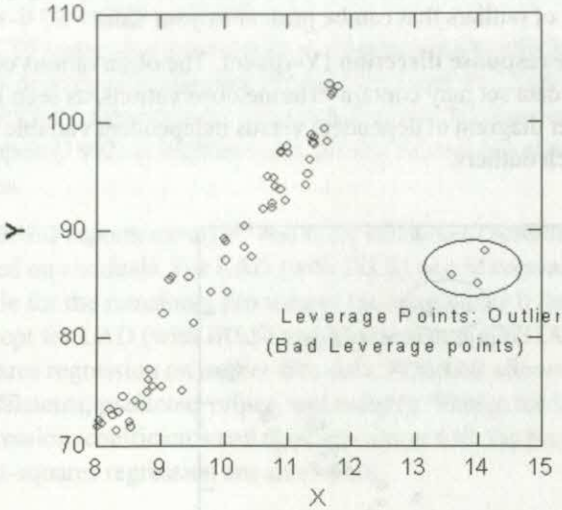
Robust regression analysis provides an alternative to the ordinary least-squares regression analysis when the fundamental assumptions on which the latter is based, are not fulfilled by the data. Assumptions such as normality, independence of observations, homoskedasticity, etc. may not be fulfilled in many practical situations. Also, there is a possibility of outliers being present in the data. Outliers are extreme observations that may be caused by large errors, typographical errors in recording the data, or by other similar reasons. When outliers are present in the data or there are violations of the assumption of normality of residuals, the estimates of regression coefficients and their standard errors are affected. Consequently, the predictive accuracy of the fitted model decreases. Often a transformation will not eliminate the influence of outliers that bias the prediction. Robust regression is mainly used for detecting outliers and for getting stable regression coefficients in the presence of outliers (Rousseeuw and Leroy, 1987).

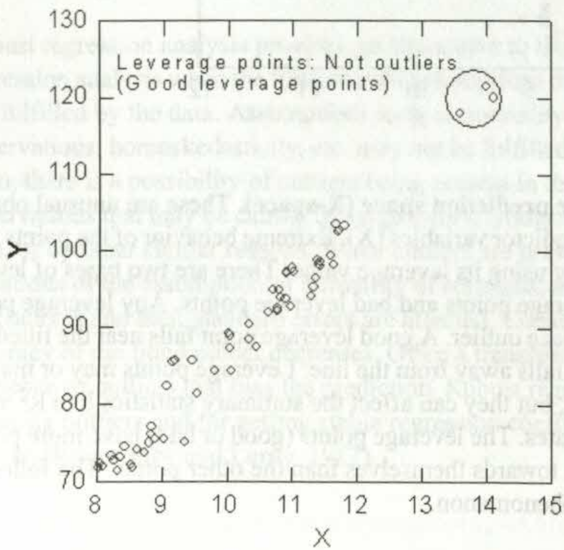
Outliers in data set may be classified into one of the three following tyes:There are mainly three types of outliers that can be present in your data:

- **Outliers in the response direction (Y-space)**. The observations on the response variable in the data set may contain extreme observations, as seen in the figure below. A scatter diagram of dependent versus independent variable may reveal the presence of such outliers.

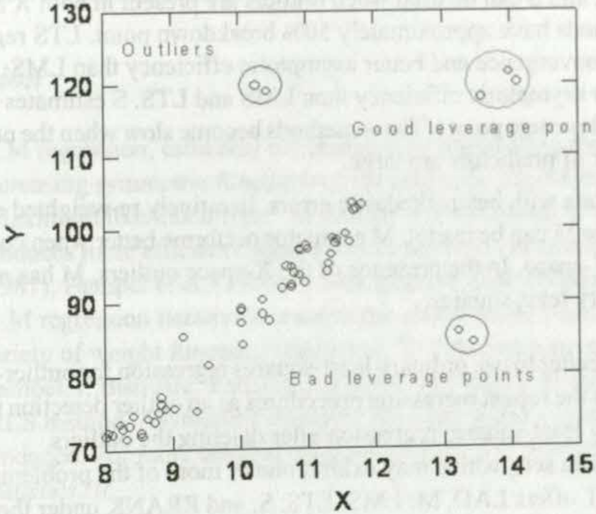


- **Outliers in the prediction space (X-space).** These are unusual observations on one or more predictor variables (X). Extreme behavior of the points in X-space can be measured by using its leverage value. There are two types of leverage points, viz., good leverage points and bad leverage points. Any leverage point (good or bad) is an X-space outlier. A good leverage point falls near the fitted line, but a bad leverage point falls away from the line. Leverage points may or may not affect the regression line, but they can affect the summary statistics like R^2 and standard errors of estimates. The leverage points (good or bad) have more power to pull the regression line towards themselves than the other points. The following figures illustrate this phenomenon.


Leverage Points: Outlier
(Bad Leverage points)
X


Leverage points: Not outliers
(Good leverage points)
X

■ **Outliers in X-space and Y-space.** These are the outlying observations in both X and Y-spaces. If the data contain outliers as well as leverage points, then the regression line as well as statistics like R^2, standard error, etc. are badly affected. The outliers as well as the bad and the good leverage points can be visualized by using a scatter-plot as follows:



Before proceeding to regression analysis, it is always informative to visualize data through a scatter-plot of dependent variable and independent variable. It may help to choose an appropriate regression model for the analysis. However it is difficult to visualize the outliers in higher dimensions.

To fit a regression model in situations mentioned above, several robust regression procedures are available. A choice of an appropriate procedure is made depending upon the two important properties of a robust estimator; breakdown point and efficiency. The breakdown point is defined as the fraction of outlying observations that may cause the estimator to take an arbitrarily large aberrant value. Usually, a robust estimator with a high breakdown point is preferred. The highest breakdown point that an estimator can achieve is 50% (Rousseeuw and Leroy, 1987). The efficiency of a robust estimator can be thought of as a ratio of residual mean square of OLS and residual mean square of the robust procedure. A good robust estimator should give efficiency value close to unity.

The following are some of the points that may be helpful in choosing an appropriate regression in the given situations:

- When Y-space outliers are present, or errors are skewed or non-normal; LAD regression may provide better estimates. LAD regression is sensitive to X-space outliers. Rank regression is a nonparametric method and is based on very few assumptions, especially distribution assumptions.
- LTS, LMS, and S can be used when outliers are present in both X and Y spaces. These methods have approximately 50% breakdown point. LTS regression has smoother convergence and better asymptotic efficiency than LMS. S estimates have higher asymptotic efficiency than LMS and LTS. S estimates are highly resistant to leverage points. These methods become slow when the number of cases and number of predictors are large.
- In case of data with heteroskedastic errors, iteratively re-weighted estimation methods like M can be useful. M estimator performs better when contamination is mainly in Y-space. In the presence of the X-space outliers, M has no advantage over ordinary least-squares.

Generally, it is better to use ordinary least-squares regression for outlier-free data. One can make use of the robust regression procedures as an outlier detection technique, and use the ordinary least-squares regression after deleting the outliers.

To analyze data sets, which may exhibit one or more of the problems discussed above, SYSTAT offers LAD, M, LMS, LTS, S, and RRANK under the Robust Regression feature. The following discussion gives some technical details of the various procedures.

Least Absolute Deviations (LAD) Regression

LAD method was introduced by Boscovich in 1757 (Birkes and Dodge, 1993). It estimates regression coefficients by minimizing the sum of absolute values of residuals. LAD estimates, on average, are less influenced by outliers than least-squares estimates. LAD regression in SYSTAT uses two methods for estimation, viz., Iteratively Reweighted Least-Squares (IRLS) and modified simplex (Birkes and Dodge, 1993). The IRLS method minimizes the sum of weighted squared residuals iteratively, where weights are reciprocal to absolute residuals at each iteration. The Simplex algorithm calculates the sum of absolute deviations at some point and searches for the direction in which this value is further minimized. It stops at the point where the sum of absolute deviations is minimum. It produces estimates of the regression coefficients

that are robust to the outliers present in the Y-direction. The LAD regression method is not robust to leverage points; the breakdown point of LAD regression is 0%. However, if the data contain less serious outliers but the underlying distribution is non-normal and heavy-tailed (for example, double exponential or Cauchy or t-distribution), LAD method should be preferred to fit a linear regression model. When Simplex method is opted, LAD regression reports the standard errors and Wald's confidence intervals for estimated regression coefficients.

M Regression

In M regression, estimates are obtained by minimizing the sum of less rapidly increasing symmetric functions of the residuals. The M in M estimation stands for 'Maximum likelihood type'. M estimator is not robust to leverage points but it produces more efficient estimates when the errors are normally distributed. See Huber (1981), Hampel et al. (1986) or Montgomery et al. (2006) for detailed discussions.

M regression iteratively weights the observations based on values of residuals. A variety of weight functions (known as Ψ functions) are used viz., Huber, Trim, Hampel, *t*, Bisquare, Ramsay, Andrews, and Tukey. M estimates are computed using IRLS method. Asymptotic standard error and Wald's confidence interval are also produced. For more detailed information refer to "Nonlinear Models" on page 261 in *Statistics III*.

Least Median Squares (LMS) Regression

Rousseeuw (1984) proposed the least median of squares (LMS) method of estimation. This method minimizes the median of the squared residuals. LMS estimator is robust with respect to the outliers in Y-spaces as well as X-spaces. LMS estimator is a high breakdown point estimator.

LMS regression implements two search procedures: Quick Search and Exhaustive Search. If the number of predictors and the number of cases are large, then the Quick Search method is used. This method is computationally expensive for large data.

Least Trimmed Squares (LTS) Regression

Least trimmed squares (LTS) regression is based on the subset of size h (out of n cases) whose least-squares fit possesses the smallest sum of squared residuals. The size of the

subset can be set to any value between n/2 and n. Initially, this method was proposed by Rousseeuw (1984) to yield a highly robust regression estimator, with a high breakdown value (n-h)/n but it is computationally expensive. Rousseeuw and Van Driessen (2000) developed a computationally efficient algorithm called FAST-LTS. The basic idea in this algorithm is an inequality involving order statistics and sums of squared residuals. The intercept adjustment technique is also used in this algorithm. For small data sets, FAST-LTS typically finds the exact LTS, whereas for larger data sets it gives more accurate results than the previous LTS algorithm and is faster by orders of magnitude. LTS (Li, 2005) estimator is robust with respect to outliers in Y-space as well as in X-space. In fact, it has the highest breakdown point of 50%.

Scale (S) Regression

The S estimate was proposed by Rousseeuw and Yohai (1984). S estimates are highly resistant to leverage points. This estimator is based on a scale estimator of residuals, and so it is called S estimator. These estimates are obtained by minimizing the sum of less rapidly increasing Tukey's bi-weighted symmetric functions of the residuals which give weights to the cases (observations).

SYSTAT implements an improved algorithm for S estimation proposed by Ruppert (1992). The efficiency of S estimate is controled by two parameters, viz. breakdown point and C value. C value is the constant in Tukey's function.

Rank Regression

Rank regression is a nonparametric regression method. Like all non-parametric methods, it is based on very few assumptions, especially distributional assumptions. It is useful when data have outliers. Rank regression calculates nonparametric estimates of the parameters of a linear model using the method of weighted median. It is expected to perform reasonably well for almost all possible distributions of the errors. It is based on the idea of using ranks of residuals instead of the observed residuals themselves.

Asymptotic Standard Errors, Confidence Intervals and Robust R^2

ROBREG computes asymptotic standard errors (ASE) for LAD and M regression using central differencing finite approximation of the Hessian matrix. S regression offers three options for the asymptotic standard error of estimated regression coefficients

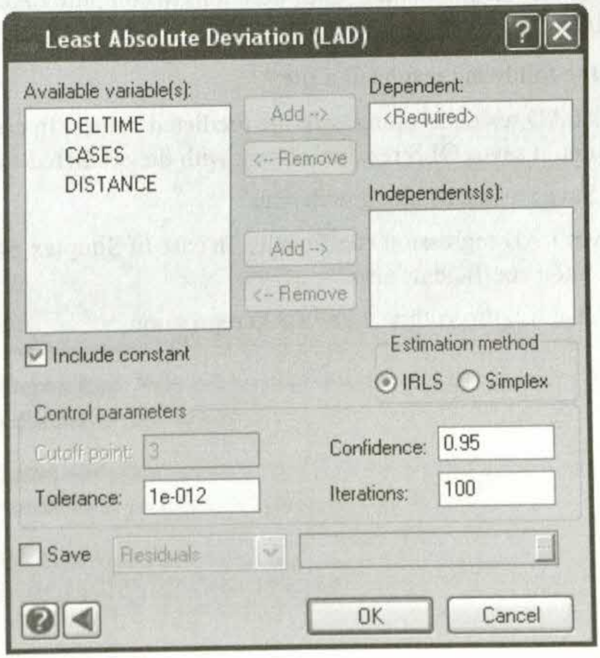
with the COV = H1 or H2 or H3. For computational details refer to Computation section of this chapter. The standard errors are computed from the diagonal elements of the estimated asymptotic variance-covariance matrix. For rank regression, variance-covariance matrix of estimated coefficients is $\tau^2(X'X)^{-1}$ where τ plays a role similar to σ in ordinary least-squares regression. Finally, ROBREG offers robust R^2 to asses the adequacy of the model.(Rousseeuw and Van Driessen, 2000 and Rousseeuw and Yohai, 1984).

Robust Regression in SYSTAT

Least Absolute Deviation (LAD) Regression Dialog Box

To open the Least Absolute Deviation (LAD) regression dialog box, from the menus choose:

Analyze
  Regression
    Robust
      Least Absolute Deviation (LAD)...



Dependent. Select the variable to be predicted. The dependent variable should be continuous and numeric.

Independent(s). Select one or more independent variables. The independent variable(s) should be continuous and numeric.

Include constant. Includes the constant in the regression equation. Deselect this option to fit a model without intercept.

Estimation method. Specify a method to compute LAD estimates:

- **IRLS.** Select this option to compute LAD estimates using IRLS method. It is the default option.
- **Simplex.** Select this option to compute LAD estimates using Simplex algorithm.

Control parameters. You can specify the following control parameters:

- **Cutoff point**. Specify the cutoff point for detecting outliers. The default is 3.
- **Confidence**. Specify a confidence level to compute the confidence intervals for regression coefficients. The default is 0.95.
- **Tolerance**. Specify the tolerance for LAD estimates. The default is 1e-12.
- **Iterations**. Specify the maximum number of iterations to compute LAD estimates. The default is 100.

Save. You can save the following results to a file:

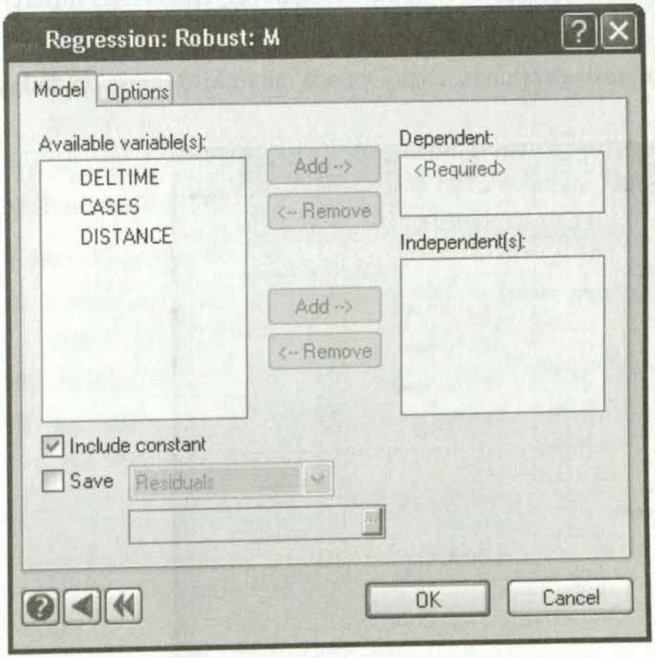
- **Residuals**. Saves LAD residuals along with the predicted values. In case of Simplex estimation, it saves OLS residuals along with the predicted values.
- **Residuals/data**. Saves residuals along with data.
- **Coefficients**. Saves LAD regression coefficients. In case of Simplex estimation, it saves OLS regression coefficients also.
- **Weights**. Saves case weights estimated by LAD regression.

M Regression Dialog Box

M Regression: Model

To open the M regression dialog box, from the menus choose:

Analyze
  Regression
    Robust
      M...



Dependent. Select the variable to be predicted. The dependent variable should be continuous and numeric.

Independent(s). Select one or more independent variables. The independent variable(s) should be continuous and numeric.

Include constant. Includes the constant in the regression equation. Deselect this option to fit a model without intercept.

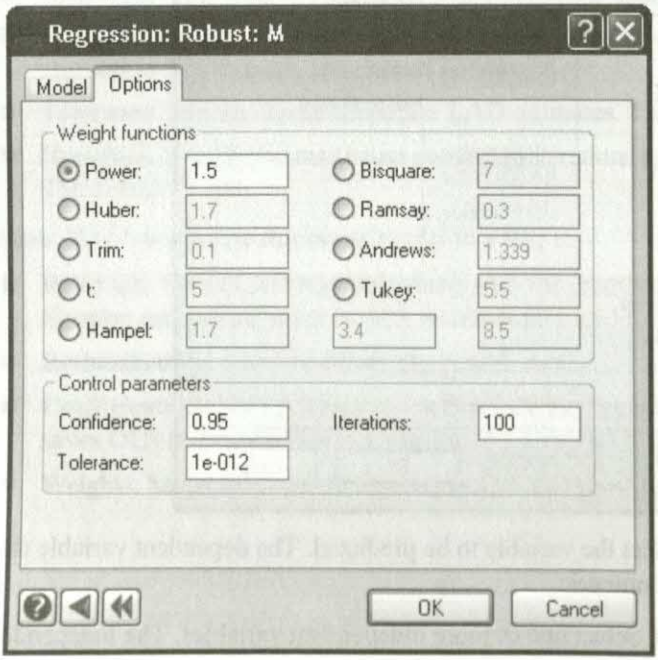
Save. You can save the following results to a file:

- **Residuals**. Saves M residuals along with the predicted values.
- **Residuals/Data**. Saves residuals along with data.
- **Coefficients**. Saves M regression coefficients.
- **Weights**. Saves case weights estimated by M regression.

M Regression: Options

To downweight the influence of extreme cases, M regression uses different weight functions.

To specify M regression options, click Options tab in M Regression dialog box.



Weight functions. The available weight functions are as follows:

- **Power**. The sum of the n^{th} power of absolute values of residuals. The default is 1.5.
- **Huber**. The sum of MAD (Median Absolute Deviations) standardized residuals weighted by Huber. The default is 1.7.

- **Trim**. Trims n proportion of the residuals (those with the largest absolute values) and minimizes the sum of squares of the remaining residuals. The default is 0.1.
- **t**. A t distribution with *df* degrees of freedom. The default *df* is 5.
- **Hampel**. The sum of MAD standardized residuals weighted by Hampel. The default is (1.7, 3.5, 8.5).
- **Bisquare**. The sum of MAD standardized residuals weighted by Bisquare. The default is 7.
- **Ramsay**. The sum of MAD standardized residuals weighted by Ramsay. The default is 0.3.
- **Andrews**. The sum of MAD standardized residuals weighted by Andrews. The default is 1.339.
- **Tukey**. The sum of MAD standardized residuals weighted by Tukey. The default is 5.5.

The parameters for Huber, Hampel, Bisquare, Ramsay, Andrews, and Tukey are defined in MAD units (Median Absolute Deviations from the median of the residuals).

Control parameters. You can specify the following control parameters:

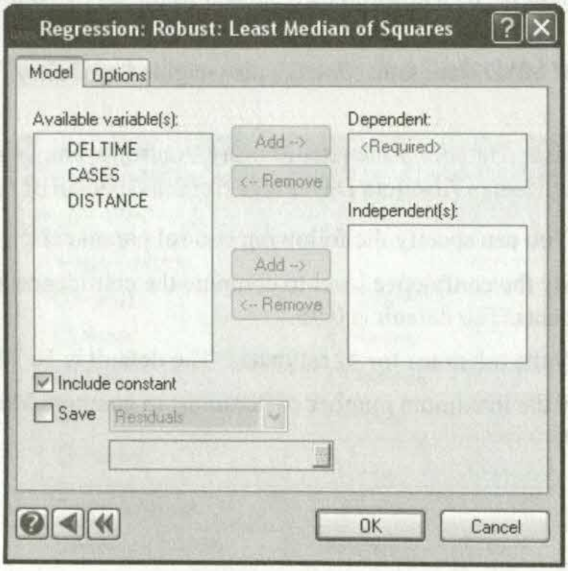
- **Confidence**. Specify the confidence level to compute the confidence intervals for regression coefficients. The default is 0.95.
- **Tolerance**. Specify the tolerance for M estimates. The default is 1e-12.
- **Iterations**. Specify the maximum number of iterations to compute M estimates. The default is 100.

Least Median of Squares (LMS) Regression Dialog Box

LMS Regression: Model

To open the Least Median of Squares (LMS) regression dialog box, from the menus choose:

Analyze
  Regression
    Robust
      Least Median of Squares (LMS)...



Dependent. Select the variable to be predicted. The dependent variable should be a continuous and numeric variable.

Independent(s). Select one or more independent variables. The independent variable should be continuous and numeric.

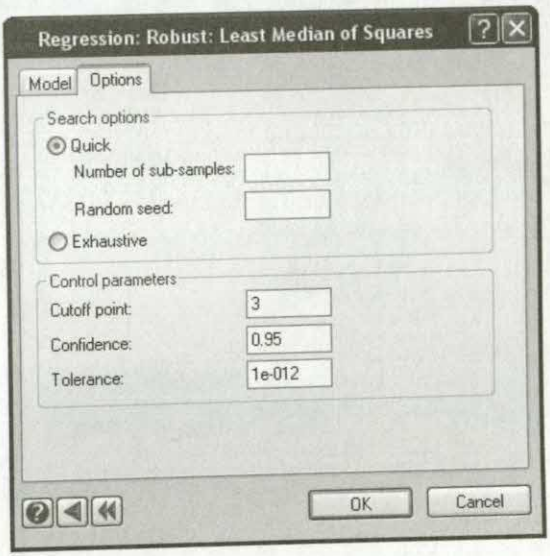
Include constant. Includes the constant in the regression equation. Deselect this option to fit model without intercept.

Save. You can save the following results to a file:

- **Residuals**. Saves LMS residuals and OLS residuals along with the predicted values.
- **Residuals/Data**. Saves residuals along with data.
- **Coefficients**. Saves LMS and OLS regression coefficients.
- **Weights**. Saves case weights estimated by LMS regression.

LMS Regression: Options

To specify LMS regression options, click Options tab in Least Median of Squares dialog box.



Search options. Choose the method for computing LMS regression estimates. Available methods are:

- **Quick**. Draws a specified number of sub-samples and reports the regression coefficients which has the minimum sum of median of squared residuals.
  - **Number of sub-samples**. Specify the number of sub-samples to compute the estimates. If the number of sub-samples is not specified, SYSTAT computes the appropriate number.

- **Random seed**. Specify the random seed to initialize the random number generator. The default random seed is generated by the system.

- **Exhaustive**. Draws all ${}^{n}C_{p}$ sub-samples and reports the regression coefficients, which have a minimum sum of median of squared residuals.

If the number of observations (n) and the number of parameters (p) are such that ${}^{n}C_{p} < 500$, then SYSTAT uses the exhaustive search method; if ${}^{n}C_{p} > 10000$, the quick search method is used, irrespective of the search method selected. If $500 < {}^{n}C_{p} < 10000$ you can choose any one of the methods.

Control parameters. You can specify the following control parameters:

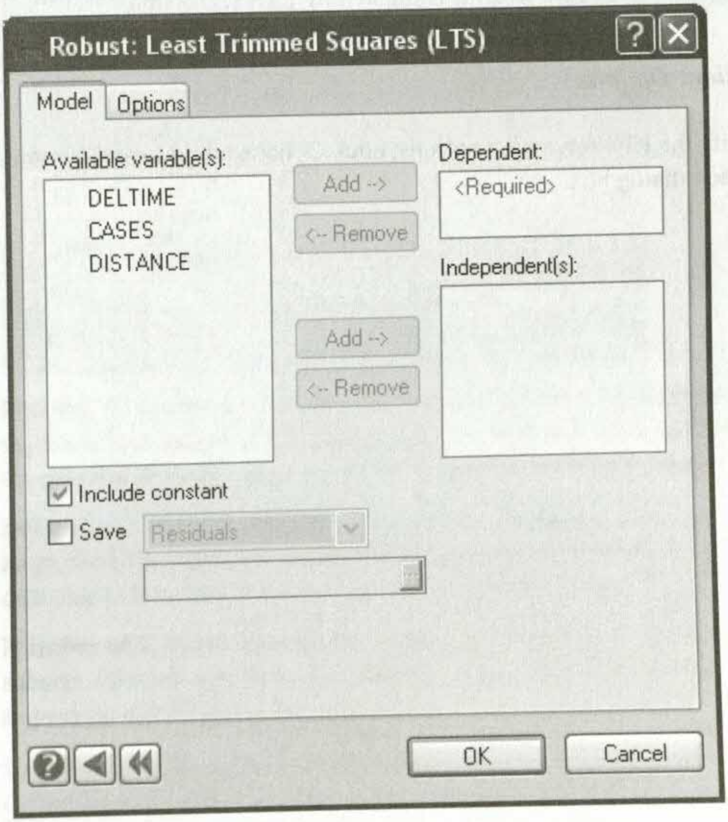
- **Cutoff point**. Specify the cutoff point for detecting outliers. The default is 3.
- **Confidence**. Specify a confidence level to compute the confidence intervals for regression coefficients. The default is 0.95.
- **Tolerance**. Specify the tolerance for LMS estimates. The default is 1e-12.

Least Trimmed Squares (LTS) Regression Dialog Box

LTS Regression: Model

To open the Least Trimmed Squares (LTS) regression dialog box, from the menus choose:

Analyze
  Regression
    Robust
      Least Trimmed Squares (LTS)...



Dependent. Select the variable to be predicted. The dependent variable should be continuous and numeric.

Independent(s). Select one or more independent variables. The independent variable(s) should be continuous and numeric.

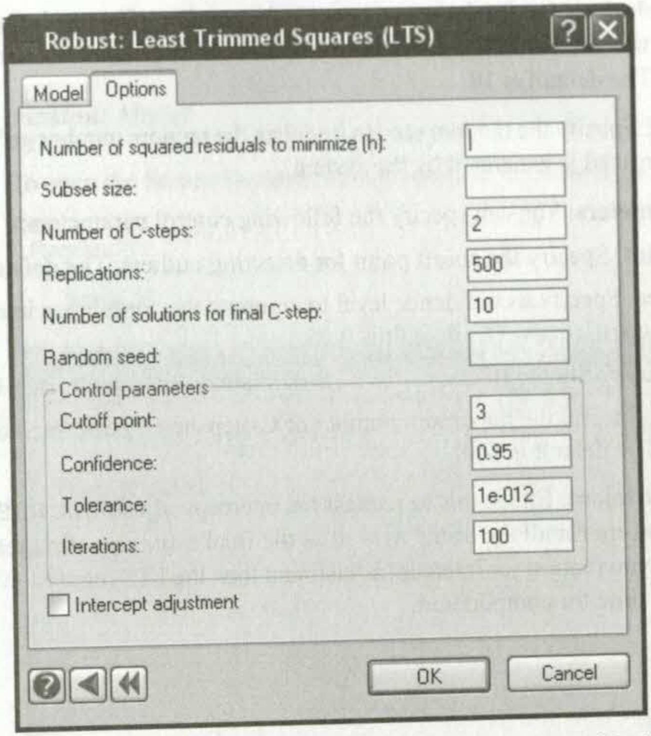
Include constant. Includes the constant in the regression equation. Deselect this option to fit a model without intercept.

Save. You can save the following results to a file:

- **Residuals**. Saves LTS residuals and OLS residuals along with the predicted values.
- **Residuals/Data**. Saves residuals along with data.
- **Coefficients**. Saves LTS and OLS regression coefficients.
- **Weights**. Saves case weights estimated by LTS regression.

LTS Regression: Options

To specify the LTS regression options, click Options tab in Least Trimmed Squares Regression dialog box.



Number of squared residuals to minimize (h). Specify the number of squared residuals to minimize (h). If you do not specify h, it is taken as (3n+p+1)/4, where n is the number of cases and p is number of parameters to be estimated.

Subsets size. Specify the size of the subsets. The default is min (n, 300). If data size is large, the LTS regression divides the data into a few subsets of specified size, and if the data size is less than double of the size of subsets, it uses full data without partition.

Number of C-steps. Specify the number of C-steps to be carried out on all the initial subsets. Given a subset, C-step obtains a new subset which gives a better estimate of regression coefficients. The default is 2.

Replications. Specifies the number of initial subsets (samples of size p each) to be drawn from the data. C-steps are performed on each subset to obtain a better subset. Input value should be less than ${}^{n}C_{p}$. The default is 500.

Number of solutions for final C-step. Specify the number of best solutions to be selected from the initial solutions. For final solutions, C-steps are carried out until convergence. The default is 10.

Random seed. Specify the random seed to initialize the random number generator. The default random seed is generated by the system.

Control parameters. You can specify the following control parameters:

- **Cutoff point**. Specify the cutoff point for detecting outliers. The default is 3.
- **Confidence**. Specify a confidence level to compute the confidence intervals for regression coefficients. The default is 0.95.
- **Tolerance**. Specify the tolerance for LTS estimates. The default is 1e-12.
- **Iterations**. Specify the maximum number of C-steps to be performed on final solutions. The default is 100.

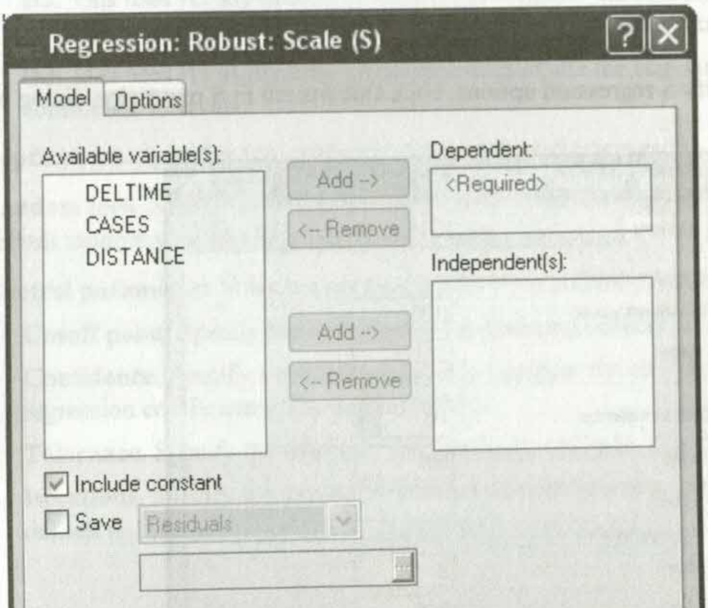
Intercept adjustment. Check this to request for intercept adjustment. ROBREG adjusts the intercept for all the initial as well as the final estimates of regression coefficients. If you request for intercept adjustment then the LTS regression procedure may take more time for computation.

S Regression Dialog Box

S Regression: Model

To open the Scale (S) regression dialog box, from the menus choose:

Analyze
  Regression
    Robust
      Scale (S)...



Dependent. Select the variable to be predicted. The dependent variable should be continuous and numeric.

Independent(s). Select one or more independent variables. The independent variable(s) should be continuous and numeric.

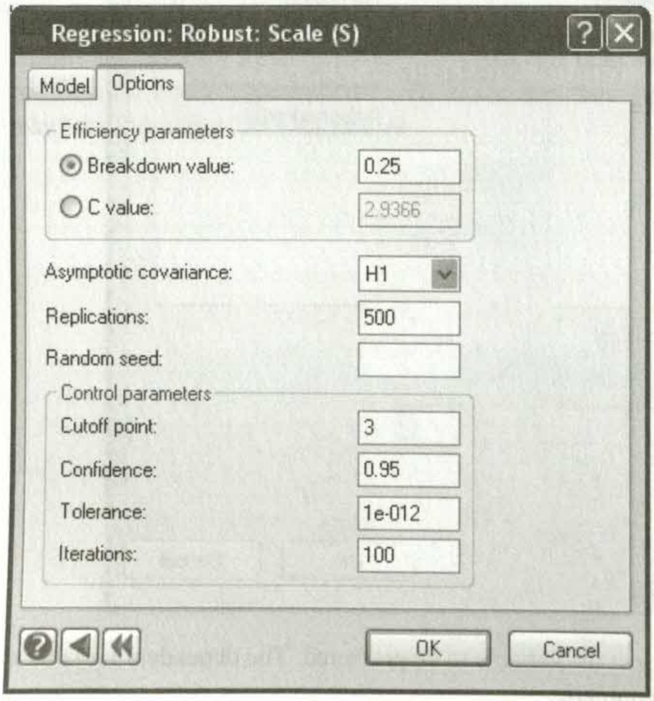
Include constant. Includes the constant in the regression equation. Deselect this option to fit a model without intercept.

Save. You can save the following results to a file:

- **Residuals**. Saves S residuals and OLS residuals along with the predicted values.
- **Residuals/Data**. Saves residuals along with data.
- **Coefficients**. Saves S and OLS regression coefficients.
- **Weights**. Saves case weights estimated by S regression.

S Regression: Options

To specify S regression options, click Options tab in S regression dialog box.



Efficiency parameters. You can specify either of the following parameters to control the efficiency of S estimate:

- **Breakdown value**. Specify the breakdown point. If you know the percentage of data contamination, specify it in terms of fraction. The value must lie between 0 and 0.5. The default is 0.25.
- **C value**. Specify any positive number. The default is 2.9366.

Asymptotic covariance. Choose the asymptotic method to compute variance-covariance matrix of estimated coefficients. The following options are available:

- **H1**. This uses H1 asymptotic covariance to compute the standard error and confidence intervals. It is the default option.
- **H2**. This uses H2 asymptotic covariance to compute the standard error and confidence intervals.
- **H3**. This uses H3 asymptotic covariance to compute the standard error and confidence intervals.

Replications. Specifies the number of sub-samples (of size p each). The default is 500.

Random seed. Specify the random seed to initialize the random number generator. The default random seed is generated by the system.

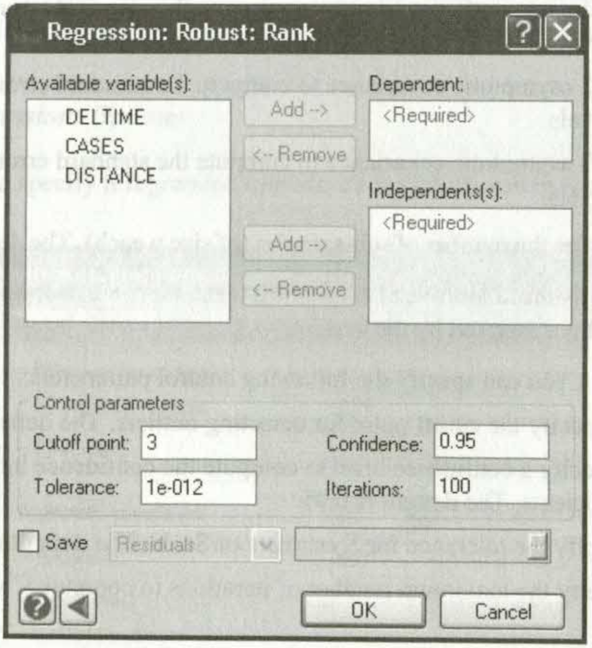
Control parameters.You can specify the following control parameters:

- **Cutoff point**. Specify the cutoff point for detecting outliers. The default is 3.
- **Confidence**. Specify a confidence level to compute the confidence intervals for regression coefficients. The default is 0.95.
- **Tolerance**. Specify the tolerance for S estimate of Scale. The default is 1e-12.
- **Iterations**. Specify the maximum number of iterations to compute C value. The default is 100.

Rank Regression Dialog Box

To open the Rank regression dialog box, from the menus choose:

Analyze
Regression
Robust
Rank...



Dependent. Select the variable to be predicted. The dependent variable should be a continuous and numeric variable.

Independent(s). Select one or more independent variables. The independent variable(s) should be continuous and numeric.

Control parameters. You can specify the following control parameters:

- **Cutoff point**. Specify the cutoff point for detecting outliers. The default is 3.
- **Confidence**. Specify a confidence level to compute the confidence intervals for regression coefficients. The default is 0.95.
- **Tolerance**. Define the convergence criterion for estimation. The default is 1e-12.

- **Iterations**. Specify the maximum number of iterations to compute the estimates. The default is 100.

Save. You can save the following results to a file:

- **Residuals.** Saves rank regression residuals and OLS residuals along with the predicted values.
- **Residuals/Data.** Saves residuals along with data.
- **Coefficients**. Saves rank and OLS regression coefficients.
- **Weights**. Saves case weights estimated by rank regression.

Using Commands

```
USE filename
ROBREG
  MODEL dependent = CONSTANT + independent(s)

  LAD/IRLS or SIMPLEX
  or
  M/POWER=n or HUBER=n or TRIM=n or T=n or BISQUARE=n or RAMSAY=n
  or ANDREWS=n or TUKEY=n or HAMPEL=n1, n2, n3
  or
  LMS/ QS or ES, NSAMPLE = n
  or
  LTS/H=n1 SSUBS=n2, NCSTEP=n3, NREP=n4, NBSOL=n5,
  INTADJUST
  or
  S/NREP=n2, BDP=n3 or C=n3 COV = H1 or H2 or H3
  or
  RRANK

  SAVE filename/RESIDUALS DATA WEIGHTS COEFFICIENTS
  ESTIMATE/TOLERANCE=k1 CONFI=k2 CUTOFF=k ITER=n
```

Usage Considerations

Types of data. ROBREG uses rectangular data only.

Print options. For LAD with Simplex method, LMS, LTS, and S regression, PLENGTH MEDIUM displays scale estimate, Robust R^2, robust estimated coefficients, number of outliers, and ordinary least-squares regression for outlier-free data. PLENGTH LONG adds ANOVA in ordinary least-squares regression for outlier-free data. For LAD with IRLS method, M, and Rank regression, output is standard for all PLENGTH options.

Quick Graphs. ROBREG produces a scatter plot of residuals against estimated values of dependent variable.

Saving files. Saves the coefficients, residuals, and predicted values to a data file.

BY groups. ROBREG analyzes data by groups.

Case frequencies. ROBREG uses the FREQUENCY variable to duplicate cases. This inflates the degrees of freedom to the sum of the number of frequencies.

Case weights. ROBREG uses the value of any WEIGHT variable to weight each case..

Examples

Example 1
Outliers in Y-space

The *PHONECAL* data set, which comes from the Belgian Statistical Survey and was analyzed by Rousseeuw and Leroy (1987), describes the number of international phone calls from Belgium in years 1950-1973. Apparently, there is a heavy contamination caused by a different measurement systems in the years 1964-1969 and parts of the years 1963 and 1970---instead of the number of phone calls, the total number of minutes of these calls was reported. Because of this new measurement system, a scatter-plot of calls against year indicates higher values for the years 1963 to 1970. Except these contaminated data, the relation between year and calls appears to be linear. This is an example of Y-space outliers. The goal is to investigate a linear relationship between phone calls and years. We use the LTS regression method to illustrate the robust regression procedure.

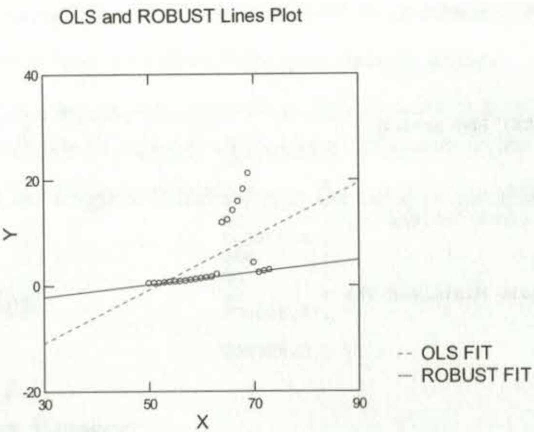
LTS Regression

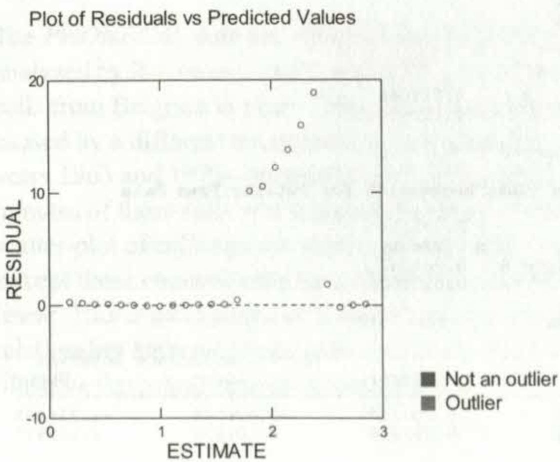
The input is:

```
USE PHONECAL
ROBREG
  MODEL Y = CONSTANT + X
  LTS / H = 12 NCSTEP = 2 NREP = 500 NBSOL = 10
  ESTIMATE
```

The output is:

```
Dependent Variable :  Y
No. of Cases       : 24
No. of Regressors  :  1

Least Trimmed Squares (LTS) Regression

Size of Subset                                 :       24
Number of C-Steps                              :        2
Maximum Number of Replications                 :      500
Number of Solutions for Final C-Steps          :       10
Cutoff Point                                   : 3.000000
Confidence Level                               :      95%
Intercept Adjustment                           :       NO
Number of Squared Residuals Minimized (h)      :       12
Breakdown Value                                : 0.500000
Robust R-square                                : 0.984960

LTS Parameter Estimates

Effect   | Coefficient
---------+------------
CONSTANT | -5.620766
X        |  0.116121

Scale Estimates

Scale (LTS)               : 0.111046
Scale (Weighted)          : 0.112905

Number of Outliers Detected :        8

Ordinary Least Squares (OLS) Regression for Outlier Free Data

Multiple R                  : 0.993136
Squared Multiple R          : 0.986319
Adjusted Squared Multiple R : 0.985342
Standard Error              : 0.097152

OLS Parameter Estimates

         |                                     95.00% Confidence Interval
Effect   | Coefficient  Standard Error           Lower             Upper
---------+-------------------------------------------------------------------
CONSTANT |  -5.170929        0.203107        -5.606549         -4.735309
X        |   0.108583        0.003418         0.101252          0.115913

Analysis of Variance

Source     |       SS  df   Mean Squares       F-ratio    p-value
-----------+-------------------------------------------------------
Regression | 9.526501   1       9.526501   1009.330324   0.000000
Residual   | 0.132138  14       0.009438

Durbin-Watson D Statistic    : 0.862158
First Order Autocorrelation  : 0.364990
```





SYSTAT displays the number of detected outliers in the data, using the fitted robust regression line. You can view the weights to see which observations are outliers (weight = 0). Here eight outliers has been detected.

After deleting the detected outliers from the data set, SYSTAT performs the ordinary least-squares regression; one can get more detailed output using the REGRESS module and using weights given by the robust regression procedure.

Example 2
Outliers in X-space

The data in the *DELTIME* data file are from Montgomery et al. (2006). In this problem, an industrial engineer is interested in predicting the amount of time required by a route driver to service the vending machines in an outlet. He observed that the two important variables that affect the delivery time (*DELTIME*) are the number of cases of product stocked (*CASES*) and the distance walked by the route driver (*DISTANCE*). Twenty-five observations were collected. It was observed that there are two outliers present in the data, namely observation numbers 9 and 22. These points influence the least-squares line of fit and in turn, you get inefficient estimates. One way is to delete such points, and fit the least-squares line. However, a more appropriate approach is to use robust regression methods. Here, we illustrate the M regression method to fit a linear regression model to the above data.

M Regression

The input is:

```
USE DELTIME
ROBREG
  M
  MODEL DELTIME = CONSTANT + CASES + DISTANCE
  ESTIMATE
```

The output is:

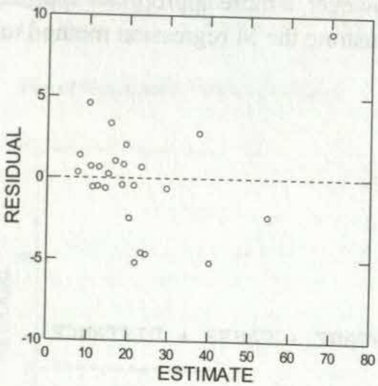
```
Dependent Variable : DELTIME
No. of Cases       :      25
No. of Regressors  :       2

M Regression

POWER is used as a Weight Function to downweight the influence of outliers.

25 cases have positive psi-weights.
The Average Psi-weight : 0.97803

Raw R-square (1-Residual/Total)                    : 0.986877
Mean corrected R-square (1-Residual/Corrected)     : 0.958459
R-square(Observed vs Predicted)                    : 0.959529

M Parameter Estimates
                                                        95.00% Confidence Interval
Effect   | Coefficient       ASE    Parameter/ASE            Lower
---------+-------------------------------------------------------------
CONSTANT |    3.026015  1.164716         2.598071         0.610542
CASES    |    1.533705  0.193209         7.938042         1.133013
DISTANCE |    0.014552  0.004024         3.616709         0.006208
```

```
M Parameter Estimates (contd...)
          |
Effect    |        Upper
----------+-------------
CONSTANT  |     5.441489
CASES     |     1.934396
DISTANCE  |     0.022896
```

Sum of Squares and Mean Squares

```
Source          |
                |          SS   df   Mean Squares
----------------+--------------------------------
Regression      | 18070.335096   3    6023.445032
Residual        |   240.293904  22      10.922450
Total           | 18310.629000  25
Mean corrected  |  5784.542600  24
```

Plot of Residuals vs Predicted Values



M regression has used POWER as a default weight function to downweight the outliers.

Example 3
Outliers in X-space and Y-space

We use the Scottish hill races data from Chatterjee and Hadi (2006) to illustrate the use of robust regression when outliers are present in both Y and X spaces. The data set consists of the record times (in seconds) of 35 Scottish Hill races in 1984 along with two explanatory variables, *DISTANCE* of race (in miles) and *CLIMB* (in feet). It was found that the records for race numbers 7 and 18 are outliers in the response direction while the points 11, 33, and 35 are outliers in X-space. Here we illustrate the S regression method to fit a linear regression model to the above data.

S Regression

The input is:

```
USE HILLRACE
ROBREG
  S / COV = H2
  MODEL TIME = CONSTANT + DISTANCE + CLIMB
  ESTIMATE
```

The output is:

```
Dependent Variable : TIME
No. of Cases       :  35
No. of Regressors  :  2
```

Scale(S) Regression

```
Size of Subset                     :        3
Maximum Number of Replications     :      300
Cutoff Point                       : 3.000000
Confidence Level                   :      95%
Asymptotic Covariance              :       H2
Breakdown Value                    : 0.250000
C Value                            : 2.937015

Robust R-square                    : 0.966123
```

S Parameter Estimates

Effect	Coefficient	Standard Error	95.00% Confidence Interval Lower	Upper
CONSTANT	-481.170995	97.413795	-679.596404	-282.745586
DISTANCE	398.519977	10.852259	376.414650	420.625305
CLIMB	0.384872	0.052866	0.277187	0.492557

```
Scale Estimate                 : 335.347573
Number of Outliers Detected :         3
```

Ordinary Least Squares (OLS) Regression for Outlier Free Data

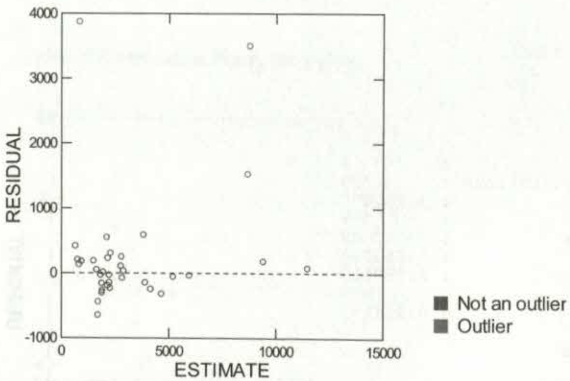
```
Multiple R                     :   0.993265
Squared Multiple R             :   0.986575
Adjusted Squared Multiple R :      0.985649
Standard Error                 : 283.124223
```

OLS Parameter Estimates

Effect	Coefficient	Standard Error	95.00% Confidence Interval Lower	Upper
CONSTANT	-493.795133	93.157700	-684.324026	-303.266241
DISTANCE	398.090821	11.919035	373.713658	422.467984
CLIMB	0.395414	0.053331	0.286340	0.504488

```
Analysis of Variance

Source     |               SS    df      Mean Squares       F-ratio    p-value
-----------+--------------------------------------------------------
Regression |  1.708272E+008     2   85413593.510873   1065.547804   0.000000
Residual   | 2324620.447003    29     80159.325759

Durbin-Watson D Statistic    :  2.158950
First Order Autocorrelation : -0.093168
```

Plot of Residuals vs Predicted Values



SYSTAT reports default or specified options used in the robust regression algorithm. Here S regression uses all the default values and the specified asymptotic covariance H2.

Now let us compare the regression estimates reported by different regression techniques.

```
Estimator    | CONSTANT   DISTANCE   CLIMB
-------------+---------------------------
OLS          | -539.483    373.073   0.663
LAD-IRLS     | -549.503    396.137   0.443
LAD-simplex  | -560.713    400.888   0.427
M-Hampel     | -535.607    399.143   0.423
M-Bisquare   | -487.303    398.282   0.390
LMS          |  -42.269    288.150   0.496
LTS          | -489.397    400.847   0.370
S            | -481.171    398.520   0.385
RANK         | -578.970    393.963   0.493
```

You can observe that the ordinary least-squares estimates are different (particularly the regression coefficient of *CLIMB*) from other robust estimates because of the presence of outliers in the data. The estimates reported by LAD and M methods may not be reliable for this data set becuase of the X-space outliers. LMS, LTS, or S regression estimates may be preferred in this case.

Computation

Algorithms

For LMS regression, the algorithm developed by Rousseeuw (1984) is employed.

For LTS regression, the FAST-LTS algorithm developed by Rousseeuw and Van Driessen (2000) is employed.

For S regression, the algorithm proposed by Ruppert (1992) is employed.

For scale regression three estimators of the asymptotic variance-covariance matrix of the robust estimator are available in SYSTAT to compute standard errors of estimated robust regression coefficients as follows:

$$\text{H1:}\quad k^2 \frac{\frac{1}{n-p}\sum_{i=1}^{n}\left(\rho'\left(\frac{r_i}{S}\right)\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\rho''\left(\frac{r_i}{S}\right)\right)^2}(X'X)^{-1} \qquad \text{H2:}\quad k\frac{\frac{1}{n-p}\sum_{i=1}^{n}\left(\rho'\left(\frac{r_i}{S}\right)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\rho''\left(\frac{r_i}{S}\right)}W^{-1}$$

$$\text{H3:}\quad \frac{1}{k(n-p)}\sum_{i=1}^{n}\left(\rho'\left(\frac{r_i}{S}\right)\right)^2 W^{-1}(X'X)W^{-1}$$

where $k = 1 + \frac{p}{n}\frac{\mathrm{var}(\rho'')}{(E(\rho''))^2}$ is a correction factor and $Wjk = \sum_{i=1}^{n}\rho''(r_i)x_{ij}x_{ik}$

For detailed information refer Huber, 1981, page 173.

For Rank regression, the method of weighted median is employed to compute the nonparametric estimates of regression coefficients .

In Rank regression τ (TAU) is computed as follows:

Suppose E_i is the i^{th} element of residuals vector E. Now consider

$$A_{ij} = \frac{E_i + E_j}{2} \quad for\ 1 \le i \le j \le n$$

Arrange these N= n(n+1)/2 pairwise averages in increasing order.

$$A(1) \le A(2) \le \ldots \le A(N)$$

Let

$$a = \frac{n(n+1)}{4} \qquad b = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$k_1 = [1/2 + a - 1.645*b] \qquad k_2 = [1/2 + a + 1.645*b]$$

$$f = \frac{n}{n-p-1}$$

$$\tau = \frac{f\sqrt{n}\left(A(k_1) - A(k_2)\right)}{3.29} = TAU$$

Missing Data

Cases with missing data are deleted in this procedure.

References

Birkes, D. and Dodge, Y. (1993). *Alternative methods of regression.* New York: John Wiley & Sons.

Chatterjee, S. and Hadi A. S. (2006). *Regression analysis by example*, 4th ed., Hoboken, N.J.: Wiley-Interscience.

*Coakley, C. W. and Hettamansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, 88, 872-880.

*Draper , N. R. and Smith, H. (1998). *Applied regression analysis*. 3rd ed. New York: John Wiley & Sons.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: John Wiley & Sons.

Huber, P. J. (1981). *Robust statistics.* New York: John Wiley & Sons.

Li, L. M. (2005). An algorithm for computing exact least-trimmed squares estimate of simple linear regression with constraints, *Computational Statistics & Data Analysis*, 48, 717-734.

Montgomery, D. C., Peck E. A., and Vining G. G. (2006). *Introduction to linear regression analysis*, 4th ed. Hoboken, N.J.: Wiley-Interscience.

Pena, D., and Yohai, V. (1999). A Fast Procedure for Outlier Diagnostics in Large Regression Problems, *Journal of the American Statistical Association*, 94, 434-445.

*Richard, W. and Holland, P. W. (1977). Two robust alternatives to least-squares regression, *Journal of the American Statistical Association*, 72, 828-833.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, Vol. 79, 871-880.

Rousseeuw P. J. and Leroy A. M. (1987). *Robust regression and outlier detection.* New York: John Wiley & Sons.

Rousseeuw, P. J. and Van Driessen, K. (2000). Computing LTS Regression for Large Data Sets. (2000). Data Mining and Knowledge Discovery, Volume 12, Number 1, January 2006, pp. 29-45(17).

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and Time Series Analysis*, edited by Franke, J., Hardle, W., and Martin, D. Lecture Notes in Statistics No. 26, New York: Springer-Verlag, pp. 256-272.

Ruppert, D. (1992). Computing S estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*,1, 253-270.

*Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, 15, 642-656.

(* indicates additional references.)

Chapter

9

Set and Canonical Correlations

Jacob Cohen and Leland Wilkinson

SETCOR computes set correlations (Cohen, 1982) and canonical correlations (Hotelling, 1935, 1936). Although it is based on algorithms developed initially for the mainframe program CORSET (Cohen and Nee, 1983) and subsequently for the PC program SETCORAN (Eber and Cohen, 1987), the SYSTAT program has a completely new source code, incorporating all the recent corrections in the statistical tests.

Finally, SETCOR also computes the Stewart and Love (1968) canonical redundancy index and rotates canonical variates.

Resampling procedures are available in this feature.

Statistical Background

Set correlation (SC) is a realization of the multivariate general linear model and is therefore a natural generalization of simple and multiple correlation. In its standard form, it generalizes bivariate and multiple regression to their multivariate analogue. The standard univariate and multivariate methods provided by the SYSTAT GLM module (for example, multivariate analysis of variance and covariance, discriminant function analysis) may be viewed as special cases of SC. SC thus provides a single general framework for the study of association. In contrast to canonical correlation, it yields a partitioning of variance in terms of the original variables, rather than their canonical transformations.

Sets

The building blocks of SC are sets of variables, which may be categorical or quantitative. They may also comprise interactions or products of measured variables. If they are nonlinearly related to each other, they should be transformed prior to analysis to avoid misleading conclusions. The same assumptions underlying ANOVA, linear regression, and other linear models are appropriate to SC.

Partialing

By partialing a set A from a set B (residualizing B by A), we produce a new set $B|A$ whose variables have zero correlations with the set A variables. (The notation $B.A$, used in some of the cited papers, is equivalent to the notation $B|A$ here.) This device has several uses in data analysis, including the statistical adjustment for irrelevant or spurious sources of variance or covariance (as in the analysis of covariance), the representation of curvilinear components and of interactions (Cohen, 1978), and the representation of contrasts among means.

In multiple regression and correlation (MRC), the use of sets and partialing applies to the right side of the equation, to the independent variables, which is where the multiplicity lies. The dependent variable y is a single variable. SC is a generalization of MRC such that a set of dependent variables Y may be related to a set X, either of which may be a partialed set. Given that virtually any information may be expressed as a set of variables, SC offers the possibility of a flexible general data-analysis method.

The basic reference for SC is Cohen (1982), reprinted in Cohen and Cohen (1983, Appendix 4), referred to hereafter as CSC. Cohen and Nee (1984) give estimators for two measures of association (shrunken $R^2_{Y,X}$ and $T^2_{Y,X}$), and Cohen (1988b, Chapter 10) provides a full treatment of power analysis in SC. van den Burg and Lewis (1988) describe the properties of the association measures and provide formal proofs. The various devices for the representation of information as sets of variables are described and illustrated in detail in Cohen and Cohen (1983, Chapters 4–9, 11), referred to hereafter as C&C. This chapter focuses on the "nuts and bolts" of the method and illustrates its chief features, as represented in the input and output of SETCOR.

Notation

In what follows, the symbols Y_B and X_B represent basic sets: set Y_B may be a set of dependent variables *Y*, or a set of dependent variables *Y* from which another set Y_P has been partialed, represented as $Y|Y_P$. Similarly, set X_B may be a set of independent variables *X*, or a set of independent variables *X* from which another set X_P has been partialed, $X|X_P$. (The term *basic* replaces the term *generic* used in CSC.) All references to the sets *Y* and *X* in the expressions that follow are to be understood to mean Y_B and X_B, the "left-hand" and "right-hand" sets, whether or not either has been partialed. Where *Y* and Y_B or *X* and X_B must be distinguished, this will be done so in the notation.

Measures of Association Between Sets

It is desirable that a measure of association between sets be a natural generalization of multiple R^2, bounded by 0 and 1, invariant over full-rank linear transformation (rotation) of either or both sets, and symmetrical (that is, $R^2_{Y,X} = R^2_{X,Y}$). Of the measures of multivariate association that have been proposed (Cramer and Nicewander, 1979), three have been found to be particularly useful: multivariate $R^2_{Y,X}$ and the symmetric ($T^2_{Y,X}$) and asymmetric ($P^2{}_{Y,X}$) squared trace correlations.

$R^2{}_{Y,X}$ Proportion of Generalized Variance

Using determinants of correlation matrices,

$$R^2_{Y,X} = 1 - \frac{|\mathbf{R}_{YX}|}{|\mathbf{R}_{YY}| \cdot |\mathbf{R}_{XX}|} \quad ,$$

where

- $\mathbf{R}_{YX}$ is the full correlation matrix of the Y_B and X_B variables,
- $\mathbf{R}_{YY}$ is the matrix of correlations among the variables of set Y_B, and
- $\mathbf{R}_{XX}$ is the matrix of correlations among the variables of set X_B.

This equation also holds when variance-covariance (**S**) or sums of squares and cross-products (**C**) matrices replace the correlation matrices.

$R^2_{Y,X}$ may also be written as a function of the q squared canonical correlations (C^2) where $q = min\ (k_Y, k_X)$, the number of variables in the smaller of the two basic sets:

$$R^2_{Y,X} = 1 - (1 - C_1^2)(1 - C_2^2)\ldots(1 - C_q^2)$$

$R^2_{Y,X}$ is a generalization of the simple bivariate $r^2_{y,x}$ and of multiple R^2 and is properly interpreted as the proportion of the generalized variance (multivariance) of set Y_B accounted for by set X_B (or vice versa, because like all product-moment correlation coefficients, it is symmetrical). The generalized variance is the generalization of the univariate concept of variance to a set of variables and is defined here as the determinant of the covariance matrix of the variables in the set. You can interpret proportions of generalized variance much as you interpret proportions of variance of a single variable. $R^2_{Y,X}$ does not vary with changes in location or scale of the variables, with nonsingular transformations of the variables within each set (for example, orthogonal or oblique rotations), or with different single degree-of-freedom codings of nominal scales. $R^2_{Y,X}$ makes possible a multiplicative decomposition in terms of squared (multiple) partial (but not semipartial) correlations. See CSC and van den Burg and Lewis (1988) for the justification of these statements and a discussion of these and other properties of $R^2_{Y,X}$.

$T^2_{Y,X}$ and $P^2_{Y,X}$ Proportions of Additive Variance

Two other useful measures of multivariate association are based on the trace of the variance-covariance matrix, $\mathbf{M}_{YX} = \mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$, where Y and X are again taken as basic. It can be shown that the trace of this matrix,

$$tr(\mathbf{M}_{YX}) = \sum_{i=1}^{q} C_i^2$$

is the sum of the q squared canonical correlations. $T^2_{Y,X}$, the symmetric squared trace correlation, is the trace divided by q, or the mean of the q squared canonical correlations.

$$T^2_{Y,X} = \frac{tr(\mathbf{M}_{YX})}{q} = \frac{\sum_{i=1}^{q} C_i^2}{q}$$

A space may be defined by a set of variables, and any nonsingular linear transformation (for example, rotation) of these variables defines the same space. Assume that $k_Y < k_X$ and consider any orthogonalizing transformation of the (basic) Y variables. Find the multiple R^2's of each of the orthogonalized Y variables with set X_B; their sum equals $tr(\mathbf{M}_{YX})$, so the mean of these multiple R^2's is $T^2_{Y,X}$. This symmetric squared trace correlation also has a proportion of variance interpretation, but unlike $R^2_{Y,X}$, the definition of variance is that of additive (or total) variance, the sum of the unit variances of the smaller set; that is, q. $T^2_{Y,X}$ provides an additive decomposition into squared semipartial (but not partial) correlations. It may, however, decrease when a variable is added to the lesser of k_Y and k_X (CSC; van den Burg and Lewis, 1988).

$P^2_{Y,X}$ is the trace divided by k_Y, the number of dependent variables, and is therefore asymmetric. When $k_Y > k_X$, its maximum is k_X / k_Y. It shares with $R^2_{Y,X}$ and multiple R^2 the property that the addition of a variable to either X or Y cannot result in a decrease. When $k_Y \leq k_X$ (the usual case), $P^2_{Y,X} = T^2_{Y,X}$. In a comprehensive analysis of their properties, van den Burg and Lewis (1988) argue that $P^2_{Y,X}$ (rather than $T^2_{Y,X}$) is a direct generalization of multiple R^2.

Interpretations

The varied uses of partialing (residualization), made familiar by MRC, make possible a functional analysis of data in terms of research factors as defined above. The basic set X_B, made up of $X|X_P$, may be used in the following ways:

- The statistical control of the research factor(s) in X_P when X is to be related to Y_B. If the model to be tested posits an independent effect of X on Y_B, then $X|X_P$ holds X_P constant; without X_P partialed, the effect found for X may be a spurious consequence of the association of X_P with both X and Y_B. In the analysis of covariance (univariate and multivariate), partialing the covariates also has the effect of reducing the error variance, and thus increases power.
- The representation of interactions of any order for research factors of any kind. For example, the $U \times V$ interaction set is constructed as $X|X_P$, where X is UV, the set of $k_U k_V$ product variables that result from multiplying each of the variables in the

research factor U by each of the variables in the research factor V, and X_P is $U + V$, the $k_{YU} + k_{YV}$ variables of the combined U and V research factors (C&C, Chapter 8).

- The representation of curve components in polynomial (curvilinear) regression. For example, for the cubic component of a variable v, the set X is v^3 and the set X_P is made up of v and v^2 (C&C, Chapter 6).
- The representation of a particular contrast within a set of means of the categories of a nominal scale. Here, X contains a single suitably coded variable and X_P contains the remaining variables carrying other contrasts (C&C, Chapter 5).
- The "purification" of a variable to its "uniqueness," as when X is made up of one subtest of a battery of intercorrelated measures and X_P contains the remaining subtests. Examples of X are the Digit Symbol subtest of the Wechsler Adult Intelligence Scale or the Schizophrenia scale score of the Minnesota Multiphasic Personality Inventory, with X_P in each instance being the respective remaining subtest/scale scores.
- The use of missing data as positive information. Here, X represents a research factor for which some subjects, having no data, are assigned an arbitrary constant (usually the mean), and X_P is a single binary variable coded 1 for the cases with missing data and 0 for those with data present (C&C, Chapter 7).

In SC, the partialing devices described above for the set X_P may equally be employed in the Y_B set as $Y|Y_P$. Thus, you may control a dependent variable for age, sex, and socioeconomic status, or represent curve components, interactions, "missingness", or uniqueness of a dependent variable or a set of dependent variables. (See CSC for examples).

Types of Association between Sets

Given the option of partialing, there are five types of association possible in SC:

	Set Y		Set X
Whole	set Y	with	set X
Partial	set $Y\|Y_P$	with	set $X\|X_P$ (where X_P=Y_P)
Y semipartial	set $Y\|Y_P$	with	set X
X semipartial	set Y	with	set $X\|X_P$
Bipartial	set $Y\|Y_P$	with	set $X\|X_P$

Formulas for the covariance matrices required for the computation of $R^2_{Y,X}$ and $T^2_{Y,X}$ for the five types of association are given in CSC, Table 1. Following an SC analysis, further analytic detail is provided by the output for MRC analyses for each basic *y* variable on the set of basic *x* variables, *y* and *x* being single variables in their respective sets. Thus, it is for the individual variables, partialed or whole depending on the type of association, that the regression and correlation results are provided. The information provided in the output for these individual basic variables (betas, multiple R^2's) facilitates the interpretation of the SC results of the X_B and Y_B sets that they constitute.

Testing the Null Hypothesis

Throughout this section, *X* and *Y* are to be understood as basic. For purposes of testing the hypothesis of no association between sets *X* and *Y*, we treat *Y* as the dependent variable set, *X* as independent, and employ the fixed model. The Wilks's (1932) likelihood ratio Λ is the ratio of the determinant of the error covariance matrix **E** to the determinant of the sum of the error and hypothesis **H** covariance matrices,

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$$

where **H** is the variance-covariance accounted for in the *Y* variables by *X*, where

$$\mathbf{H} = S_{YX}S_{XX}^{-1}S_{XY}$$

The definition of **E** depends on whether the test is to employ Model 1 or Model 2 error. Model 1 error is defined as

$$\mathbf{E_1} = S_{YY} - S_{Y,XX\mathrm{p}}R_{XX\mathrm{p}}^{-1}S_{XX\mathrm{p},Y}$$

that is, the residual covariance matrix when the covariance associated with sets *X* and X_p has been removed.

The Model 2 error is employed when there exists a set *G*, made up of variables in neither *X* nor X_p, that can be used to account for additional variance in S_{YY} and thus reduce **E** below $\mathbf{E_1}$ in the interest of unbiasedness and increased statistical power. This occurs when, with multiple research factors, the analyst wishes to use "pure" error, for example, the within-cell variance in a factorial design. In this case, the error-reducing set *G* is made up of the variables comprising the research factors ("main effects") and interactions other than the factor or interaction under test, as is done traditionally in MANOVA and MANCOVA factorial designs.

$$\mathbf{E_2} = \mathrm{S}_{YY} - \mathrm{S}_{Y,XX\mathrm{p}G}\mathrm{R}_{XX\mathrm{p}G}{}^{-1}\mathrm{S}_{XX\mathrm{p}G,Y}$$

In whole and Y semipartial association, where X_p does not exist, it is dropped from $\mathbf{E_1}$ and $\mathbf{E_2}$. Formulas for the **H** and **E** matrices for the five types of association are given in CSC, Table 2. (See Algorithms for corrections to the CSC formulas).

When Model 1 error (no set G) is used, for the whole, partial, and Y semipartial types of association, it can be shown that

$$\Lambda = 1 - R^2_{Y,X}$$

Once Λ is determined for a sample, Rao's F test (1973) may be applied to test the null hypothesis. As adapted for SC, the test is quite general, covering all five types of association and both error models. When $k_Y = 1$, where multivariate $R^2_{Y,X}$ specializes to multiple R^2's, the Rao F test specializes to the standard null hypothesis F test for MRC. For this case, and for the case where the smaller set is made up of no more than two variables, the Rao F test is exact; otherwise, it is approximate (Cohen and Nee, 1987).

$$F = \left(\Lambda^{-\frac{1}{s}} - 1\right)\frac{v}{u}\text{, where}$$

u = numerator $df = k_Y k_X$,
v = denominator $df = ms + 1 - u/2$ where
$m = N - \max(k_{Y_\mathrm{p}}, k_{X_\mathrm{p}} + k_G) - (k_Y + k_X + 3)/2$, and

$$s = \sqrt{\frac{k_Y^2 k_X^2 - 4}{k_Y^2 + k_X^2 - 5}}$$

except that when $k_Y^2 k_X^2 = 4$, $s = 1$. For partial $R^2_{Y,X}$, set X_p = set Y_p, so $k_{X_\mathrm{p}} = k_{Y_\mathrm{p}}$ is the number of variables in the set that is being partialed, and for the whole $R^2_{Y,X}$, neither set X_p nor set Y_p exists. Further, k_{Y_p}, k_{X_p}, and k_G are 0 when the set does not exist for the type of association or error model in question. The test assumes that the variables in X are fixed and those in Y are multivariate normal, but the test is quite robust against assumption failure (Cohen and Nee, 1990; Olson, 1974).

Estimates of the Population $R^2_{Y,X}$, $T^2_{Y,X}$, and $P^2_{Y,X}$

Like all squared correlations, $R^2_{Y,X}$, $T^2_{Y,X}$, and $P^2_{Y,X}$ are positively biased. Shrunken values (almost unbiased population estimates) are given by

$$\tilde{R}^2_{Y,X} = 1 - (1 - R^2_{Y,X})\left(\frac{v+u}{v}\right)^s ,$$

$$\tilde{T}^2_{Y,X} = 1 - (1 - T^2_{Y,X})\left(\frac{w+u}{v}\right)^s \text{, and}$$

$$\tilde{P}^2_{Y,X} = \tilde{T}^2_{Y,X}\frac{k_Y}{k_X}$$

where w is the denominator *df* of the Pillai (1960) F test for

$$T^2_{Y,X}\text{: } w = q[N - k_Y - k_X - max(k_{Y\mathrm{p}}, k_{X\mathrm{p}} - 1)]$$

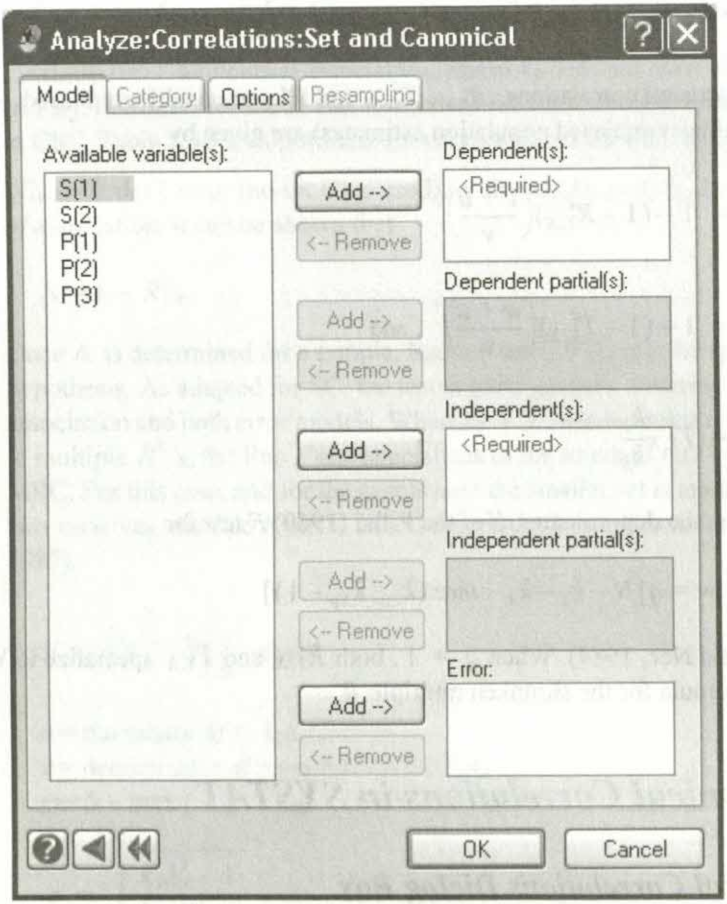
(Cohen and Nee, 1984). When $q = 1$, both $\tilde{R}^2_{Y,X}$ and $\tilde{T}^2_{Y,X}$ specialize to Wherry's (1931) formula for the shrunken multiple R^2.

Set and Canonical Correlations in SYSTAT

Set and Canonical Correlations Dialog Box

To open the Set and Canonical Correlations dialog box, from the menus choose:

Analyze
  Correlations
    Set and Canonical Correlations...



To do a SETCOR analysis, first specify a model.

Dependent(s). Enter the dependent variables you want to examine.

Dependent partial(s). You can specify the variables to be partialed out of the dependent set, which produces a new set whose variables have zero correlation with the partialed set. The partial variables are optional for dependent variables. The simple canonical correlation model does not have a partial variable list.

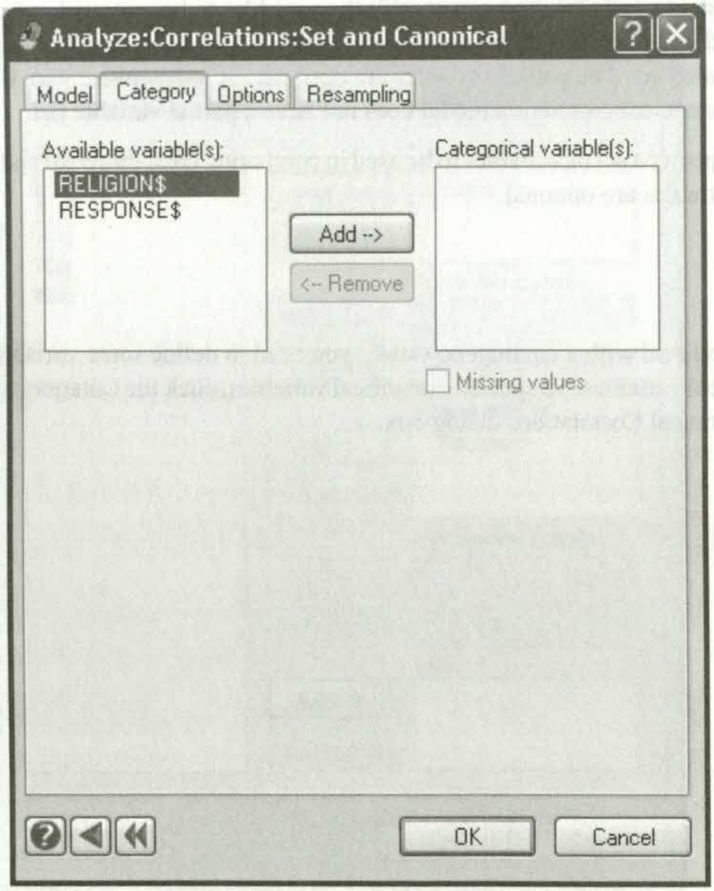
Independent(s). Select one or more continuous or categorical variables (grouping variables).

Independent partial(s). You can specify the variables to be partialed out of the independent set, which produces a new set whose variables have zero correlation with the partialed set. The partial variables are optional for independent variables. The simple canonical correlation model does not have a partial variable list.

Error. Specify a set of variables to be used in computing error terms for statistical tests. Error variables are optional.

Category

When you deal with a contingency table, you need to define some variables as categorical variables. To specify categorical variables, click the Category tab in the Set and Canonical Correlations dialog box.

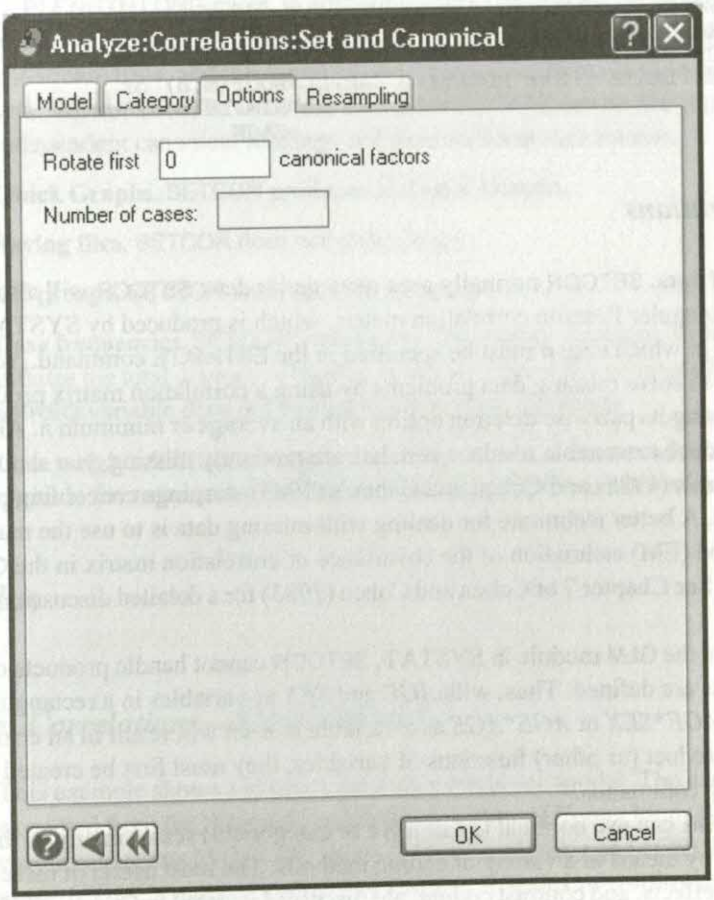


Categorical variable(s). Select one or more variables to define categories.

Missing values. Check this to consider missing values as a separate category.

Options

The Options tab controls the rotation and sample size options for estimation.



The following options can be specified:

- **Rotate first.** Enter the number of canonical factors to rotate using the varimax rotation.
- **Number of cases.** If you enter a triangular matrix (correlation), specify the number of cases from which the matrix was computed. This is required if you are using a correlation matrix instead of raw data.

Using Commands

After specifying the data with USE *filename*, continue with:

```
SETCOR
  MODEL yvarlist | ypartials = xvarlist | xpartials
  ERROR varlist
  CATEGORY varlist
  ESTIMATE / N=n ROTATE=n SAMPLE=BOOT(m,n) or
                                SIMPLE(m,n) or
                                JACK
```

Usage Considerations

Types of data. SETCOR normally uses rectangular data. SETCOR will also accept a lower triangular Pearson correlation matrix, which is produced by SYSTAT's CORR module, in which case *n* must be specified in the ESTIMATE command. You may be tempted to solve missing data problems by using a correlation matrix produced by CORR using its pairwise deletion option with an average or minimum *n*. Although this may produce reasonable results when data are randomly missing, you should consider Wilkinson's (1988) and Cohen and Cohen's (1983) warnings concerning pairwise matrices. A better technique for dealing with missing data is to use the maximum likelihood (EM) estimation of the covariance or correlation matrix in the CORR module. See Chapter 7 of Cohen and Cohen (1983) for a detailed discussion of missing data.

Unlike the GLM module in SYSTAT, SETCOR cannot handle products of variables when sets are defined. Thus, with *AGE* and *SEX* as variables in a rectangular file, naming *AGE*SEX* or *AGE*AGE* as a variable in a set will result in an error message. To use product (or other) functions of variables, they must first be created using SYSTAT commands.

SETCOR can use nominal (qualitative or categorical) scales in any of the sets it employs by means of a variety of coding methods. The most useful of these, which are dummy, effects, and contrast coding, are discussed in detail in Chapter 5 of Cohen and Cohen (1983), and their use in set correlation illustrated in Cohen (1982) and in Cohen (1988a). The CATEGORY command codes these variables.

Print options. For PLENGTH SHORT, the output gives *n*, the type of association, the variables in sets *YPARTIAL*, *XPARTIAL*, and *G* (when present), the Rao *F* (with its *df* and *p-value*), and their shrunken values, and the following results for the basic *y* and *x* variables: the within-set correlation matrices for the *y* and *x* variables, the rectangular

between-set correlation matrix, the betas for estimating each *y* variable from the *x* set (with their standard errors and *p-values*), a matrix of the intercorrelations of the estimated *y* values whose diagonal is the multiple of each *y* variable with the *x* set, and the *F* test and *p-value* for the latter.

PLENGTH LONG gives, in addition to the above results for basic *y* and *x*, the Stewart and Love redundancy index for *y* given *xb*, the canonical correlations and their Bartlett chi-square tests, and the canonical coefficients, loadings, and redundancies for both sets. When PLENGTH is LONG, the option ROTATE rotates the dependent and independent canonical loadings and the canonical correlations.

Quick Graphs. SETCOR produces no Quick Graphs.

Saving files. SETCOR does not save results.

BY groups. SETCOR analyzes data by groups.

Case frequencies. SETCOR uses the FREQ variable, if present, to duplicate cases. This inflates the total degrees of freedom to be the sum of the number of frequencies. Using a FREQ variable does not require more memory, however.

Case weights. SETCOR weights sums of squares and cross products using the WEIGHT variable for rectangular data input. It does not require extra memory.

Examples

Example 1
Canonical Correlations—Simple Model

This example shows a simple canonical correlation model. The data, which have been extracted from the National Longitudinal Survey of Young Men, 1979, includes school enrollment status (*NOTENR*, set to 1 if not enrolled), age (*AGE*), highest completed grade (*EDUC*), mother's education (*MED*), an index of reading material available in the home (*CULTURE*), and an IQ score (*IQ*) for 48 individuals.

The input is:

```
SETCOR
  USE ANXIETY
  MODEL S(1)..S(2) = P(1)..P(3)
  ESTIMATE / N=48
```

The output is:

```
SYSTAT Correlation file contains variables:

  S(1)   S(2)   P(1)   P(2)   P(3)

Whole Set Correlation Analysis (Y vs X)

Number of Cases on which Analysis is based: 48

RAO F       : 3.675
df          : 6.000 , 86.000
p-value     : 0.003

R-square   : 0.367 Shrunk R-square : 0.275
T-square   : 0.202 Shrunk T-square : 0.090
P-square   : 0.202 Shrunk P-square : 0.090

   Within Basic Set y Correlations

        ¦ S(1)    S(2)
  ------+---------------
   S(1) ¦ 1.000
   S(2) ¦ 0.303   1.000

   Within Basic Set x Correlations

        ¦ P(1)    P(2)    P(3)
  ------+-----------------------
   P(1) ¦ 1.000
   P(2) ¦ 0.304   1.000
   P(3) ¦ 0.403   0.589   1.000

   Between Basic y (col) and Basic x (row) Correlations

        ¦ S(1)     S(2)
  ------+---------------
   P(1) ¦ 0.391    0.106
   P(2) ¦ 0.298   -0.028
   P(3) ¦ 0.197    0.321

   Estimated (from x-set) y Intercorrelations (R-square on diagonal)

        ¦ S(1)    S(2)
  ------+---------------
   S(1) ¦ 0.193
   S(2) ¦ 0.002   0.175

   Significance Tests for Prediction of Each Basic y Variable

   Variable ¦ F-ratio   p-value
  ----------+-------------------
   S(1)     ¦   3.505    0.023
   S(2)     ¦   3.115    0.036

   Betas Predicting Basic y (col) from Basic x (row) Variables

        ¦   S(1)     S(2)
  ------+-----------------
   P(1) ¦   0.353   -0.001
   P(2) ¦   0.243   -0.332
   P(3) ¦  -0.088    0.517

   Standard Error of Betas

        ¦ S(1)    S(2)
  ------+---------------
   P(1) ¦ 0.149   0.150
   P(2) ¦ 0.168   0.170
   P(3) ¦ 0.175   0.177
```

t-statistic for Betas

	S(1)	S(2)
P(1)	2.374	-0.010
P(2)	1.442	-1.953
P(3)	-0.503	2.921

p-value for Betas

	S(1)	S(2)
P(1)	0.022	0.992
P(2)	0.156	0.057
P(3)	0.618	0.005

Stewart-Love Canonical Redundancy Index : 0.184

Canonical Correlations

1	2
0.511	0.377

Bartlett Test of Residual Correlations

	Chi-square	df	p-value
Correlations 1 through 2	20.087	6.000	0.003
Correlations 2 through 2	6.754	2.000	0.034

Canonical Coefficients for Dependent (y) Set

	1	2
S(1)	-0.893	0.551
S(2)	0.796	0.684

Canonical Loadings (y variable by factor correlations)

	1	2
S(1)	-0.652	0.759
S(2)	0.525	0.851

Canonical Redundancies for Dependent Set

1	2
0.092	0.092

Canonical Coefficients for Independent (x) Set

	1	2
P(1)	0.618	0.513
P(2)	0.941	-0.247
P(3)	-0.959	0.809

Canonical Loadings (x variable by factor correlations)

	1	2
P(1)	0.518	0.764
P(2)	0.564	0.385
P(3)	-0.156	0.870

```
Canonical Redundancies for Independent Set

      1       2
 ---------------
  0.053   0.071
```

Because this is a whole association, the "basics" are the original unpartialed variables. Note that there is considerable shrinkage. The overall association (*R-square* = 0.367) is nontrivial and significant (*p-value* = 0.003).

The individual regression analyses provide detail: *P*(1..3) are significantly related to *S*(1) (*multiple-R* = 0.193), with *P*(1) yielding a significant beta. They are also significantly related to *S*(2) (*multiple-R* = 0.175), with *P*(3)'s beta significant.

Example 2
Partial Set Correlation Model

This example shows a partial set correlation model. In a large-scale longitudinal study of childhood and adolescent mental health (Cohen and Brook, 1987), data were obtained on personal qualities that the subjects admired and what they thought other children admired, as well as the sex and age of the subjects. The admired qualities were organized into scales for antisocial, materialistic, and conventional values for the self and as ascribed to others. In one phase of the investigation, the researchers wanted to study the relationship between the sets of self versus others. However, several of these scales exhibited sex differences, were nonlinearly (specifically quadratically) related to age, and/or were differently related to age for the sexes. For the self-other association to be assessed free of the confounding influence of age, sex, and their interactions, it was desirable to partial those effects from the association. Accordingly, using SYSTAT commands, the variables *AGE, SEX* times *AGE* and their squares were created, which, together with *AGE* and *SEX*, constituted both the *YPARTIAL* and *XPARTIAL* sets in the partial association. The resulting rectangular data file *ADMIRE* was analyzed as follows,

The input is:

```
SETCOR
  USE ADMIRE
  MODEL ANTISO_O,MATER_O,CONVEN_O |,
        AGE,SEX,AGESQ,SEXAGE,SEXAGESQ =,
        ANTISO_S,MATER_S,CONVEN_S |,
        AGE,SEX,AGESQ,SEXAGE,SEXAGESQ
  ESTIMATE
```

The output is:

```
SYSTAT Rectangular file contains variables:

  ID$         ANTISO_S   MATER_S    CONVEN_S   ANTISO_O   MATER_O
  CONVEN_O    AGE        SEX        AGESQ      SEXAGE     SEXAGESQ

Partial Set Correlation Analysis (Y|YPAR vs X|XPAR, WHERE YPAR=XPAR)

Number of Cases on which Analysis is based: 755

Dependent Set y Partialled by these Variables
     AGE
     SEX
     AGESQ
     SEXAGE
     SEXAGESQ

Independent Set x Partialled by these Variables
     AGE
     SEX
     AGESQ
     SEXAGE
     SEXAGESQ

RAO F     : 52.169
df        :  9.000 , 1810.851
p-value   :  0.000

R-square  : 0.429 Shrunk R-square : 0.422
T-square  : 0.169 Shrunk T-square : 0.159
P-square  : 0.169 Shrunk P-square : 0.159

  Within Basic Set y Correlations

           | ANTISO_O   MATER_O   CONVEN_O
 ----------+------------------------------
  ANTISO_O |   1.000
  MATER_O  |   0.200     1.000
  CONVEN_O |  -0.417     0.105     1.000

  Within Basic Set x Correlations

           | ANTISO_S   MATER_S   CONVEN_S
 ----------+------------------------------
  ANTISO_S |   1.000
  MATER_S  |   0.206     1.000
  CONVEN_S |  -0.258     0.063     1.000

  Between Basic y (col) and Basic x (row) Correlations

           | ANTISO_O   MATER_O   CONVEN_O
 ----------+------------------------------
  ANTISO_S |   0.393     0.077    -0.066
  MATER_S  |   0.133     0.456     0.046
  CONVEN_S |  -0.111     0.120     0.351

  Estimated (from x-set) y Intercorrelations (R-square on diagonal)

           | ANTISO_O   MATER_O   CONVEN_O
 ----------+------------------------------
  ANTISO_O |   0.157
  MATER_O  |   0.052     0.216
  CONVEN_O |  -0.028     0.053     0.124
```

Significance Tests for Prediction of Each Basic y Variable

Variable	F-ratio	p-value
ANTISO_O	46.436	0.000
MATER_O	68.673	0.000
CONVEN_O	35.358	0.000

Betas Predicting Basic y (col) from Basic x (row) Variables

	ANTISO_O	MATER_O	CONVEN_O
ANTISO_S	0.377	0.009	0.022
MATER_S	0.056	0.448	0.018
CONVEN_S	-0.017	0.094	0.356

Standard Error of Betas

	ANTISO_O	MATER_O	CONVEN_O
ANTISO_S	0.036	0.034	0.036
MATER_S	0.035	0.033	0.035
CONVEN_S	0.035	0.034	0.036

t-statistic for Betas

	ANTISO_O	MATER_O	CONVEN_O
ANTISO_S	10.543	0.249	0.616
MATER_S	1.611	13.430	0.520
CONVEN_S	-0.486	2.783	9.965

p-value for Betas

	ANTISO_O	MATER_O	CONVEN_O
ANTISO_S	0.000	0.803	0.538
MATER_S	0.108	0.000	0.603
CONVEN_S	0.627	0.006	0.000

Stewart-Love Canonical Redundancy Index : 0.166

Canonical Correlations

1	2	3
0.487	0.401	0.329

Bartlett Test of Residual Correlations

	Chi-square	df	p-value
Correlations 1 through 3	418.286	9.000	0.000
Correlations 2 through 3	216.191	4.000	0.000
Correlations 3 through 3	85.145	1.000	0.000

Canonical Coefficients for Dependent (y) Set

	1	2	3
ANTISO_O	0.462	1.039	0.114
MATER_O	0.733	-0.671	0.320
CONVEN_O	0.471	0.448	-0.919

Canonical Loadings (y variable by factor correlations)

	1	2	3
ANTISO_O	0.412	0.718	0.561
MATER_O	0.875	-0.416	0.246
CONVEN_O	0.356	-0.056	-0.933

Canonical Redundancies for Dependent Set

1	2	3
0.084	0.037	0.045

Canonical Coefficients for Independent (x) Set

	1	2	3
ANTISO_S	0.392	0.986	0.077
MATER_S	0.745	-0.585	0.404
CONVEN_S	0.470	0.196	-0.910

Canonical Loadings (x variable by factor correlations)

	1	2	3
ANTISO_S	0.425	0.815	0.395
MATER_S	0.856	-0.369	0.362
CONVEN_S	0.416	-0.095	-0.904

Canonical Redundancies for Independent Set

1	2	3
0.086	0.043	0.040

The partial association is substantial (0.429), significant, and because of the large n and small x and y sets, hardly affected by shrinkage. The within and between basic x and y set correlation coefficients are all partial correlation coefficients because the basic x and y sets are respectively *X|XPARTIAL* and *Y|YPARTIAL* with XPARTIAL=YPARTIAL, and it is for these partialed variables that the multiple-regression output (betas, multiple R squares, etc.) is given.

For example, the significant beta = 0.377 for *ANTISO_S* in estimating *ANTISO_O* are for both with the variables *AGE, SEX, AGESQ, SEXAGE*, and *SEXAGSQ* partialed, and *ANTISOC_S* is further partialed by *MATER_S* and *CONVEN_S*. Note that each *_O* variable has significant betas with its paired *_S* variable. In addition, *MATER_O*'s beta for *CONVEN_S* is significant. All the partialed y variables have significant multiple R squares with the partialed x set, with *MATER_O* being the largest.

Example 3
Contingency Table Analysis

From the perspective of set correlations, a two-way contingency table displays the association between two nominal scales, each represented by a suitably coded set of variables. A nominal scale of *n* levels (categories) is coded as *n*–1 variables, and when each is partialed by the other *n*–2 variables, it carries a specific contrast or comparison, its nature depending on the type of coding employed. The major types of coding—dummy, effects, and contrast—are described in Chapter 5 of Cohen and Cohen (1983); their use in contingency table analysis is illustrated in Cohen (1982).

Zwick and Cramer (1986) compared the application of various multivariate methods in the analysis of contingency tables using a fictitious example from Marascuilo and Levin (1983), and Cohen (1988a) provides a complete set correlation analysis of this example. It is of responses by 500 men to the question "Does a woman have the right to decide whether an unwanted birth can be terminated during the first three months of pregnancy?" The response alternatives were crosstabulated with religion, resulting in the following table of frequencies:

	Protestant	**Catholic**	**Jewish**	**Other**	**Total**
Yes	76	115	41	77	309
No	64	82	8	12	166
No opinion	11	6	2	6	25
Total	151	203	51	95	500

Religion and response are represented by ordinal numbers in the data file *SURVEY3*. Religion is effects-coded as E(l), E(2), and E(3). When from each of these the other two are partialed, the resulting variable compares that group with the unweighted combination of all four groups; that is, it estimates that group's "effect." Notice how we use the FREQ command to determine the cell frequencies,

The input is:

```
SETCOR
  USE SURVEY3
  CATEGORY RELIGION$ RESPONSE$
  FREQ COUNT
  MODEL RESPONSE$=RELIGION$
  ESTIMATE
```

The output is:

```
SYSTAT Rectangular file contains variables:

  RELIGION$   RESPONSE$   COUNT

Case Frequencies Determined by Value of Variable COUNT

Categorical values encountered during processing are

  Variables             |               Levels
  ----------------------+------------------------------------------
  RELIGION$ (4 levels)  | Catholic   Jewish       Other   Protestant
  RESPONSE$ (3 levels)  | No         No_opinion   Yes

Whole Set Correlation Analysis (Y vs X)

Number of Cases on which Analysis is based: 640

RAO F       : 50.311
df          :  6.000 , 1270.000
p-value     :  0.000

R-square    : 0.347 Shrunk R-square : 0.341
T-square    : 0.182 Shrunk T-square : 0.174
P-square    : 0.182 Shrunk P-square : 0.174

  Within Basic Set y Correlations

              | RESPONSE$1   RESPONSE$2
  ------------+-------------------------
  RESPONSE$1  |    1.000
  RESPONSE$2  |   -0.569         1.000

  Within Basic Set x Correlations

              | RELIGION$1   RELIGION$2   RELIGION$3
  ------------+--------------------------------------
  RELIGION$1  |    1.000
  RELIGION$2  |   -0.123         1.000
  RELIGION$3  |   -0.269        -0.381        1.000

  Between Basic y (col) and Basic x (row) Correlations

              | RESPONSE$1   RESPONSE$2
  ------------+-------------------------
  RELIGION$1  |   -0.147         0.189
  RELIGION$2  |   -0.186         0.274
  RELIGION$3  |    0.545        -0.405

  Estimated (from x-set) y Intercorrelations (R-square on diagonal)

              | RESPONSE$1   RESPONSE$2
  ------------+-------------------------
  RESPONSE$1  |    0.298
  RESPONSE$2  |   -0.217         0.195

  Significance Tests for Prediction of Each Basic y Variable

  Variable   | F-ratio    p-value
  -----------+-------------------
  RESPONSE$  |  89.841     0.000

  Betas Predicting Basic y (col) from Basic x (row) Variables

              | RESPONSE$1   RESPONSE$2
  ------------+-------------------------
  RELIGION$1  |    0.006         0.129
  RELIGION$2  |    0.027         0.174
  RELIGION$3  |    0.557        -0.304
```

Standard Error of Betas

	RESPONSE$1	RESPONSE$2
RELIGION$1	0.036	0.019
RELIGION$2	0.037	0.020
RELIGION$3	0.038	0.020

t-statistic for Betas

	RESPONSE$1	RESPONSE$2
RELIGION$1	0.168	6.846
RELIGION$2	0.735	8.866
RELIGION$3	14.551	-15.080

p-value for Betas

	RESPONSE$1	RESPONSE$2
RELIGION$1	0.867	0.000
RELIGION$2	0.463	0.000
RELIGION$3	0.000	0.000

Stewart-Love Canonical Redundancy Index : 0.246

Canonical Correlations

1	2
0.558	0.228

Bartlett Test of Residual Correlations

	Chi-square	df	p-value
Correlations 1 through 2	271.251	6.000	0.000
Correlations 2 through 2	34.099	2.000	0.000

Canonical Coefficients for Dependent (y) Set

	1	2
RESPONSE$1	0.815	0.904
RESPONSE$2	-0.279	1.184

Canonical Loadings (y variable by factor correlations)

	1	2
RESPONSE$1	0.973	0.229
RESPONSE$2	-0.743	0.670

Canonical Redundancies for Dependent Set

1	2
0.233	0.013

Canonical Coefficients for Independent (x) Set

	1	2
RELIGION$1	0.056	0.690
RELIGION$2	0.047	1.008
RELIGION$3	-0.965	0.626

```
Canonical Loadings (x variable by factor correlations)

             |   1      2
-------------+---------------
RELIGION$1   |  0.309  0.398
RELIGION$2   |  0.408  0.684
RELIGION$3   | -0.998  0.056

Canonical Redundancies for Independent Set

   1      2
-------------
 0.131  0.011
```

The whole association is modest (0.347) but highly significant, and provides some Fisherian protection for the tests of specific hypotheses that follow. To determine where this overall association is coming from, assess the association of religion with the Yes-No contrast C(l).C(2) using y semipartial association.

To analyze the effects of the religious groups on the Yes-No contrast, we turn to the betas for E(1..3). Since these are partial regression coefficients, each reflects a comparison of its religious group with an equally weighted combination of the four groups on the Yes versus No contrast. For example, the Protestant group (Beta = 0.129) shows a greater proclivity to respond "No" (compared to "Yes") with $t = 6.846$, $df = 495$, *p-value* = 0.000. (For dealing with the implicitly coded "Other" group, see Chapter 5 of Cohen and Cohen, 1983.) Further analyses of these data using contrast functions of religious group membership and bipartial association are given in Cohen (1988a).

Computation

Algorithms

Table 2 in Cohen (1982) contains errors in two of the matrix expressions for the Y semipartial. The expression for **H** should read (in Cohen's notation)

$$\mathbf{H} = \mathbf{C}_{D \cdot C, B \cdot C} \mathbf{C}^{-1}_{B \cdot C} \mathbf{C}'_{D \cdot C, B \cdot C}$$

and in E_2, B should be replaced with B.C. The expression E_1 is correct as is. We are indebted to Charles Lewis for this correction.

Missing Data

When a rectangular data file is used in SETCOR, the program computes a Pearson correlation matrix on all the numeric variables in the file on a listwise basis. This means that if a value is missing for any variable in the file, the case is dropped and n is reduced accordingly. If the pattern of missing data makes n small, you should impute missing values by maximum likelihood (EM) in the CORR module.

References

Cohen, J. (1978). Partialed products are interactions; partialed powers are curve components. *Psychological Bulletin*, 85, 858-866.

Cohen, J. (1982). Set correlation as a general multivariate data-analytic method. *Multivariate Behavioral Research*, 17, 301–341.

Cohen, J. (1988a). Set correlation and contingency tables. *Applied Psychological Measurement*, 12, 425–434.

Cohen, J. (1988b). *Statistical power analysis for the behavioral sciences,* 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cohen, P. and Brook, J. (1987). Family factors related to the persistence of psychopathology in childhood and adolescence. *Psychiatry*, 50, 332–345.

Cohen, J. and Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences,* 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cohen, J. and Nee, J. C. M. (1983). CORSET, A Fortran IV program for set correlation analysis. *Educational and Psychological Measurement*, 43, 817–820.

Cohen, J. and Nee, J. C. M. (1984). Estimators for two measures of association for set correlation. *Educational and Psychological Measurement*, 44, 907–917.

Cohen, J. and Nee, J. C. M. (1987). A comparison of two noncentral F approximations, with applications to power analysis in set correlation. *Multivariate Behavioral Research*, 22, 483–490.

Cohen, J., and Nee, J. C. M. (1990). Robustness of Type I error and power in set correlation analysis of contingency tables. *Multivariate Behavioral Research*, 25, 341-350.

Cramer, E. M. and Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika*, 44, 43–54.

Eber, H. W. and Cohen, J. (1987). SETCORAN: A PC program to implement set correlation as a general multivariate data-analytic method. Atlanta: Psychological Resources.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, 26, 139–142.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321–377.

Marascuilo, L. A. and Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, Calif.: Brooks/Cole.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69, 894–908.

*Pedhazur, E. J. (1982). *Multiple regression in behavioral research*, 2nd ed. New York: Holt, Rinehart & Winston.

Pillai, K. C. S. (1960). *Statistical tables for tests of multivariate hypotheses*. Manila: The Statistical Institute, University of the Philippines.

Rao, C. R. (1973). *Linear statistical inference and its applications,* 2nd ed. New York: John Wiley & Sons, Inc.

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70, 160–163.

van den Burg, W. and Lewis, C. (1988). Some properties of two measures of multivariate association. *Psychometrika*, 53, 109–122.

Wherry, R. J. (1931). The mean and second moment coefficient of the multiple correlation coefficient in samples from a normal population. *Biometrika*, 22, 353–361.

Wilkinson, L. (1988). *SYSTAT. The system for statistics*. Evanston IL : Systat, Inc.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24, 471–494.

Zwick, R. and Cramer, E. M. (1986). A multivariate perspective on the analysis of categorical data. *Applied Psychological Measurement*, 10, 141–145.

(* indicates additional references.)

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321–377.

Marascuilo, L. A. and Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, Calif.: Brooks/Cole.

[illegible] (1974). [illegible] of variance [illegible]

Pedhazur, E. J. (1982). *Multiple regression in behavioral research*. 2nd ed. New York: Holt, Rinehart & Winston.

Pillai, K. C. S. (1960). *Statistical tables for tests of multivariate hypotheses*. Manila: The Statistical Institute, University of the Philippines.

Rao, C. R. (1973). *Linear statistical inference and its applications*. 2nd ed. New York: John Wiley & Sons, Inc.

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70, 160–163.

van den Burg, W. and Lewis, C. (1988). Some properties of two measures of multivariate association. *Psychometrika*, 53, 109–122.

Wherry, R. J. (1931). [illegible] coefficient of the multiple correlation [illegible]

Wilkinson, L. (1988). *SYSTAT: The system for statistics*. Evanston, Ill.: Systat, Inc.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24, 471–494.

Chapter

10

Signal Detection Analysis

Herb Stenson

The SIGNAL module provides analysis of data that are appropriate for the theory of signal detection as described by Green and Swets (1989), Egan (1975), and many others. For some interesting applications, see Swets and Pickett (1982), Swets (1986), and Kraemer (1988).

The response data to be analyzed by SIGNAL can be from 2 to 11 response categories. Thus, either binary or rating-scale data can be analyzed. An iterative technique is used to produce maximum likelihood estimates of all model parameters, including the locations of the category boundaries. Graphical displays of ROC curves are available in addition to the numerical output.

SIGNAL allows analysis based on a number of statistical models in addition to the more usual normal distribution and nonparametric models. The additional models are the logistic, negative exponential, chi-square, Poisson, and gamma distribution models. These models are useful for various types of detection tasks in which the sets of assumptions concerning the nature of the detector dictate one of these models. For a discussion of these alternative models, see Egan (1975).

The parameter estimates from any model can be saved into a SYSTAT file, as can the coordinates of any ROC curve.

Statistical Background

The theory of signal detectability (TSD) emerged after World War II as a synthesis of existing methods for representing the characteristics of a receiver or sensing device (Peterson, Birdsall, and Fox, 1954). Although its origins were in electrical engineering, the abstraction of the theory made it especially suited to analysis of

human perception in general and of any system involving detection of a weak signal against a background of noise: perception of visual and auditory signals, medical diagnosis based on signs or symptoms, robotic perception and so on. TSD is now widely used in medical research for evaluating the sensitivity and specificity of diagnostic equipment and clinicians. In radiology, for example, TSD can be used to quantify the performance of radiologists reading diagnostic X-rays when the "signal" (true diagnosis) is known from subsequent events or external criteria. See Hanley and McNeil (1982) for an example.

Detection Parameters

The signal detection theory literature abounds with various indices of the detectability of a signal and associated parameters. Because of this proliferation and the confusion that it can cause, the indices and parameters that are estimated by SIGNAL are described here in some detail. You can read more about these indices in books by Swets and Pickett (1982), Egan (1975), and Green and Swets (1989). Coombs, Dawes, and Tversky (1970) provide the best summary.

Except for the non-parametric model, the output printed by SIGNAL contains a standard set of parameters and indices of detectability for all of the models. The NPAR model is nonparametric, so there are no parameters to estimate. The only index of detection that is given for it is the area under the ROC curve obtained by joining the points on the ROC graph by straight lines. See Bamber (1975) for ways to test hypotheses about this area measure.

For every model involving a statistical distribution, SIGNAL prints the mean and standard deviation of the noise (N) distribution and the mean and standard deviation of the signal+noise ($S+N$) distribution. For compactness, let us call these *MN, SN, MS,* and *SS*, respectively. For the normal distribution model, *MN* will always be 0 and *SN* will always be unity because these two parameters are chosen as the origin and the unit of the scale for the decision axis. They are a part of the standard output because they do not take on these fixed values for all of the models.

Using these means and standard deviations, SIGNAL computes and prints three measures of the separation of the $S+N$ and N distributions for each of the models. These measures are labeled as *D-Prime*, *D Sub-A*, and *Sakitt D* in the output.

D-Prime (d') is the most common index of detectability used in detection research. It is defined as $(MS - MN)/SN$. As has been pointed out by many authors, it suffers from the lack of information about the standard deviation of the $S+N$ distribution

when this information is available. The other two measures computed by SIGNAL take this information into account.

D Sub-A (d_a) uses as a denominator the square root of the mean of the *N* and $S + N$ variances. Let us call this number *x*. Thus, you would square both *SN* and *SS*, add these squares, divide by two, and take the square root of the result in order to find *x*. Then the index *D Sub-A* is defined as $(MS - MN)/x$. This index is related to the area under the ROC curve for normal distributions and has other statistical niceties. See Simpson and Fitter (1973) for further discussion.

Sakitt D is another measure of detectability that takes into account the variances of both the *N* and $S + N$ distributions. It was proposed by Sakitt (1973). It uses as a denominator the square root of the product of *SN* and *SS*. Thus, this index is defined as $(MS - MN)/(\sqrt{SN \times SS})$. Egan (1975) proposes this as the best detection index for chi-square, gamma, and Poisson models.

In addition to these measures of the separation of the $S + N$ and *N* distributions, SIGNAL also prints in the output for each model the ratio of *SS* to *SN*. It is labeled *SD-Ratio*, which stands for *standard deviation ratio*.

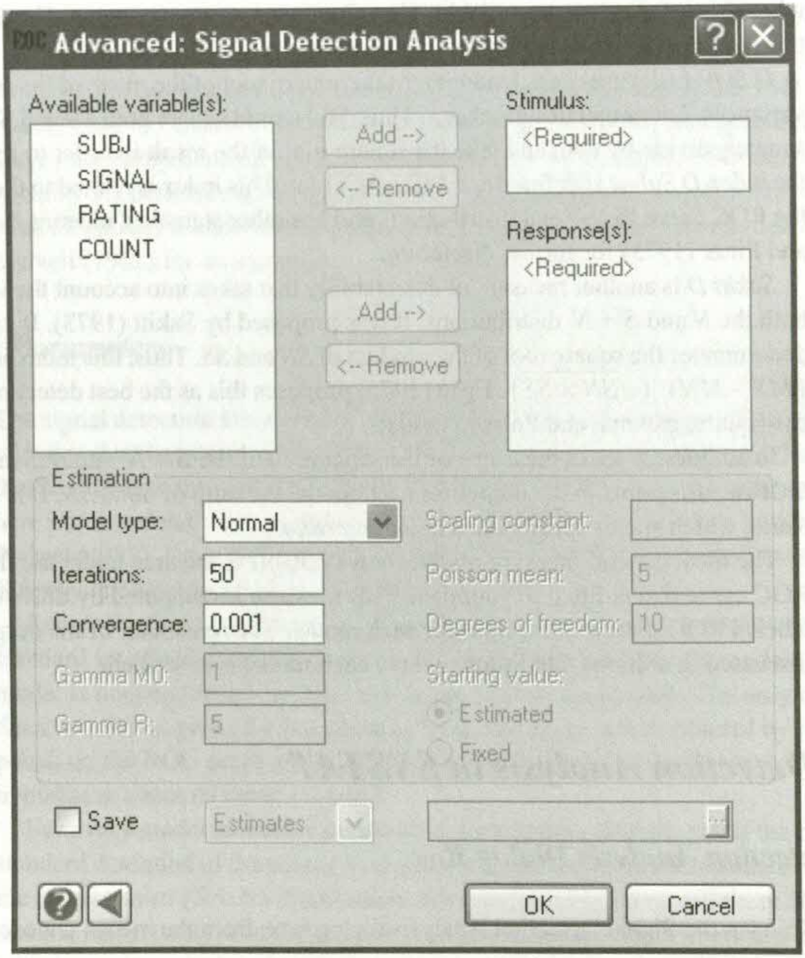
The most general measure of detection available is the area under the theoretical ROC curve that is fitted to your data. This measure is computed by SIGNAL and is labeled ROC *Area* in the output for each model. The remainder of the output is discussed in sections that follow, where each model is described.

Signal Detection Analysis in SYSTAT

Signal Detection Analysis Dialog Box

To open the Signal Detection Analysis dialog box, from the menus choose:

Advanced
  Signal Detection Analysis...



Signal detection models are computed with a model specification and estimation stage.

Stimulus. Select the variable that shows the true state of the signal on each trial for signal-detection data. The stimulus variable can only contain the numbers 0 (for noise occurrences) and 1 (for signal+noise occurrences) to indicate the stimulus state on trial. The stimulus variable remains in effect until you change it or use a different data file.

Response(s). Select the variable(s) that contain the response(s) to the stimulus by one or more detectors. The response variable can contain numbers only between –10 and +10 and there can be only 11 categories of response (for example, 0 through 10 or –5

through +5). If the input data contain decimals, they will be truncated instead of rounded. The response variable remains in effect until you change it or use a different data file.

Model type. With the exception of the nonparametric model, each model assumes that the trials on which noise alone occurred (N) and the trials on which both signal and noise occurred $(S + N)$ are samples from a particular kind of statistical distribution, with possibly different parameters for the N and $S + N$ distributions. Possible models include:

- **Chi-square.** You can think of the chi-square model as a generalization of the exponential model. To enter a fixed value for the degrees of freedom, select Fixed value and enter a positive number in the Degrees of freedom box. Alternatively you can specify a starting value for the degrees of freedom and Signal Analysis will attempt to find the best-fitting value for the degrees of freedom during iteration. To specify a starting value for degrees of freedom, select Estimation and enter a positive number in the Degrees of freedom box.
- **Exponential.** The negative exponential density function is algebraically identical to a chi-square density function with two degrees of freedom. However, its simple properties and usefulness justify treating it as a separate model.
- **Gamma.** Use the gamma model when your experiment can be described as a Poisson process and the detector uses the time required to accumulate a fixed number of rare events as the basis of the response. The length of a trial is determined by the detector as opposed to a Poisson counting observer, who counts events during fixed interval trial. You can specify both the M0 parameter and the number of events (R) that the detector is waiting to accumulate. If you do not enter a fixed value for R, the default starting value (5) will be used and the program will estimate a value of R at every iteration.
- **Logistic.** The logistic model models the noise or signal+noise distributions and is a good approximation for the normal, while being mathematically more tractable.
- **Normal.** Use this model to indicate that the noise (N) and signal+noise $(S + N)$ distributions are Gaussian.
- **Nonparametric.** A nonparametric model offers a simple way to get a quick look at your data. If you believe that the assumptions of any of the parametric models that you can use are not justifiable for your data, then use the nonparametric model.
- **Poisson.** Select a Poisson model when the detector is basing a response on a small number of countable, rare events that occur on each trial. For a Poisson model, you can specify the mean of the noise distribution. The mean of the Poisson distribution

is the average number of occurrences per trial of the event being counted. The mean for the signal+noise distribution will be estimated to give the best fit to your data.

Scaling constant. For the logistic model, the default scaling constant is $\pi/\sqrt{3}$, which is approximately 1.814. With the default value in effect, the standard deviation of the noise distribution will be 1.00. For the exponential model, the default scaling constant is unity. You can set the scaling constant to be any positive number

Iterations. This option controls the maximum number of iterations that you want to allow the program to perform in order to estimate the parameters. The default value is 50, which for most applications is more than enough. However, if you have a lot of response categories, a small value of convergence, and "difficult" data, you may need more than this for some models.

Convergence. The convergence option controls the degree of accuracy sought in the estimations. Its default value is 0.001, which means that the estimates are to be accurate to 0.001 times their values. You can set convergence to any number from 0.1 to 0.00001.

Save. Saves parameter estimates or ROC curve coordinates to a data file.

Using Commands

After selecting the data with USE *filename*, continue with:

```
SIGNAL
      MODEL responsevarlist = stimulusvar
      ESTIMATE / type  CONVERGE=d  ITERATIONS=n ,
                       M0=d  R=d  C=d  MEAN=d  DF=d
```

Type must be one of the following:

```
NORMAL              NPAR               LOGISTIC
EXPONENTIAL         CHISQUARE          POISSON
GAMMA
```

For a signal detector, the response variable list contains a single variable. Multiple detectors (for example, judges) of a single signal can be fit in a signal model.

When analyzing your data, SIGNAL computes initial estimates (starting values) for each parameter that is to be estimated by the iterative process. Typing ESTIMATE again after an analysis has finished will cause SIGNAL to use the most recent estimates of parameters as starting values for continuing the analysis, rather than computing new ones. You can, if you want, change the options for ESTIMATE when restarting the

program in this way. However, this restart procedure will not work if you specify a new MODEL or use FREQ after an analysis terminates. It also will not work if you are using the BY command. You can, however, use any of the other commands (except USE) before restarting the program.

The values of CONVERGE and ITER always revert to their default values for each use of ESTIMATE. Therefore, they must be stated explicitly each time that you use ESTIMATE if you don't want to use the default settings. The value that you use for CONVERGE is irrelevant, and therefore unnecessary, if you specify ITER=0. Because no iterations will occur, there is no accuracy of estimation to worry about. Similarly, since no iterations are used for the NPAR model, both ITER and CONVERGE are inappropriate options for this model.

The capability to restart the program using the most recent values, combined with the options for ESTIMATE, enable you to be flexible in the way that you approach the analysis of your data. You could, for example, set CONVERGE to a large value, such as 0.1, and set ITER to a small value, such as 4. This will cause the iterative process to proceed rather quickly. After a look at the output, you could restart the program by typing ESTIMATE again with a smaller value for CONVERGE, and perhaps with a different number for ITER.

Usage Considerations

Types of data. The format of input data for SIGNAL is quite flexible in order to easily accommodate data from a variety of experimental designs commonly used in signal detection studies. The program requires a SYSTAT data set containing a minimum of two numeric variables: One that shows the true state of the signal on each trial of your experiment, and the other that shows the response of a detector to that signal state. Thus, the cases in your SYSTAT file represent trials (instances of the signal or lack thereof) in a detection experiment.

If you have more than one detector responding to exactly the same sequence of stimuli, responses from the additional detectors can also be coded as variables in the SYSTAT data file. In this case, there should be only one variable that indicates the true state of the signal on each trial. You could also have more than one variable that designates the true state of the signal on each trial. For example, if each detector was exposed to a different sequence of signal states, you could have a separate variable that indicates the true state of the signal for each detector.

The example below shows how to enter data for a hypothetical experiment in which each of three detectors (HS, LB, and LW) responded on each of five trials to exactly

the same sequence of signal states. (You would, of course, have many more trials than this for a real experiment.) Imagine that a response was to be one of the numbers –2, –1, 0, 1, or 2, with a –2 indicating that the detector was sure that no signal was present on a trial, a 2 indicating that the detector was sure that a signal was present on a trial, and the other numbers indicating degrees of certainty as to whether a signal was present or not. The true state of the signal (present or not present) is coded as the variable labeled *STATE*.

```
INPUT STATE,HS,LB,LW
1  0  1  2
0 -1 -2 -1
0  0 -1  0
1  2  1  1
0 -1  0 -1
ENDINPUT
ESAVE MYFILE
```

Another way to encode the same data would be to create either a string or numeric variable to identify a detector (for example, *UNIT$*), a variable to show the true state of the stimulus (for example, *STATE*), and a third variable to indicate the response of the detector (for example, *RATING*). Then, on each line of the data set, you would enter the identifier for the detector, the state of the stimulus on the trial in question, and the response of the detector. Such a data set would then contain as many cases as there were trials times the number of detectors. You then could use the SELECT command within the SIGNAL module to identify which of the detectors you want to analyze, or you could use the BY command to analyze each detector sequentially. This would be an easy way to enter data if each detector had been exposed to a different sequence of signal states. However, this would not be an optimal way to enter data when each detector was exposed to exactly the same sequence of signal states because you are repeating the same set of numbers representing the signal states for each detector.

The availability of negative numbers as response options makes it possible to encode responses from a particular kind of signal detection task that is sometimes used. In this task, the detector (usually a human detector) is to specify first whether or not a given trial contained a signal, and then is asked to rate his or her confidence that his or her response is correct on, say, a five-point rating scale. A way to encode such data for SIGNAL would be to treat all confidence ratings on trials when the subject reported the absence of a signal as the numbers –1 through –5, and to treat the ratings on trials when the subject reported that there was a signal present as the numbers +1 through +5. A similar encoding strategy can be used when a detector reports the presence or absence of a signal, and the reaction time for the response is categorized and used as a *confidence rating*, with quick times indicating a high degree of confidence in the

response. You would encode the reaction times into categories acceptable to SIGNAL for this experimental paradigm.

SIGNAL treats the response categories as ordinal data. Thus, it makes no difference in the analysis what numbers are used, even if there are gaps in the sequence used. All that is necessary is that the higher numbers indicate a "signal-like" response and the lower ones indicate a "noise-like" response. For example, using the response categories 1, 2, and 3 would result in the same analysis as using the categories –6, 0, and +2 for the same data. Only the category labels would be affected in the program output. Notice that gaps can occur in the response category sequence either because certain numbers were not available to the detector as response options or because the detector never used one of the available options. The program obviously cannot distinguish which of these is the case.

You can specify more than one variable as a response. This allows the pooling of responses of detectors that were exposed to the same stimulus sequence, or the pooling of responses from one detector that was exposed to the same sequence of stimuli on more than one occasion. Each occasion would have to be entered into the data set as a separate variable. For example, to pool the responses from detectors HS and LW from the data set *MYFILE* (above), you would type:

```
MODEL HS,LW = STATE
```

The resulting signal detection analysis would treat all responses from these two detectors as being from the same detector. Thus, the resulting detection parameter estimates would apply to this group of detectors instead of to one of them individually.

If you have used a different coding scheme, like the one used for the data set *SWETSDTA* in the first example, where each detector has an identifier code, you could pool detectors by simply not using a SELECT or BY command when you analyze the data. SIGNAL would then treat all response entries as coming from the same detector. You could also use the SELECT command with multiple identifier variables, such as sex and age, to pool data within these subgroups. The resulting analysis would then apply to whatever group(s) you selected to pool.

Print options. The output is standard for all PLENGTH options.

Quick Graphs. SIGNAL plots the receiver operating characteristic (ROC) curve.

Saving files. If you save before you estimate, SIGNAL will save parameter estimates into a file. If you add SAVE / ROC, SIGNAL will save the ROC curve coordinates.

BY groups. SIGNAL analyzes data by groups. Your file need not be sorted on the BY variable(s).

Case frequencies. If a file contains frequencies of each type of response to each of two stimulus states, the frequencies of responses can then be used as a FREQ variable in the SIGNAL module. This can be useful if your data are already aggregated in this way or if you want to make up a table of hypothetical data to model some signal detection task.

Case weights. SIGNAL does not allow case weighting.

Examples

Example 1
Normal Distribution Model for Signal Detection

This example shows frequency data for two detectors (subjects) in a study by Swets, Tanner, and Birdsall (1961) as reported by Swets and Pickett (1982, pp. 216–219). Each of the subjects in the experiment used a six-category rating scale to indicate his or her confidence that a signal was present on each of 597 trials when the signal was present, and on 591 randomly-mixed trials on which the signal was not present. The *COUNT* variable shows the number of times a subject gave a particular rating to a given signal state. Notice that the identifier *SUBJ* is a numeric variable in this case (but would not have to be).

By far, the most common model used for signal detection analysis is the normal (Gaussian) model, in which the noise (N) distribution and the signal+noise ($S + N$) distribution are both assumed to be Gaussian density functions. These functions have the same variance in the case of binary response data, or have possibly unequal variances in the case of more than two response categories.

Here we use the data set named *SWETSDTA* that was described earlier. To perform a signal detection analysis using the normal distribution model for the first subject in the data set

the input is:

```
USE SWETSDTA
SELECT SUBJ=1
SIGNAL
MODEL RATING=SIGNAL
FREQ COUNT
ESTIMATE
```

Notice that the SELECT command is used to specify which detector to analyze. SELECT remains in effect throughout any subsequent analysis until you change the selection by using the SELECT command again (or cancel it completely by typing SELECT with nothing after it).

The FREQ command is used in the same way as it is in the rest of SYSTAT. Here it specifies the variable that shows the frequencies with which response categories were used for the two different signal states. If you were using a data set that was coded in a manner similar to *MYFILE*, you obviously would not use the FREQ command.

The output is:

```
Data for the following results were selected according to
     SELECT SUBJ=1

Case frequencies determined by value of variable COUNT

  Number of Stimulus Events (Cases) Responded to      : 1188.000
  Number of Detectors (variables) Observing an Event :    1.000
  Number of Response Categories Used                 :    6.000
  Number of Responses to Noise Events                :  591.000
  Number of Responses to Signal Events               :  597.000
  Total Number of Responses                          : 1188.000
  Number of Instances of Missing Data                :    0.000

  Response Category |  Frequency      Joint Probability     Conditional Probability
                    | Noise  Signal    Noise    Signal        Noise    Signal
  ------------------+-----------------------------------------------------------
  1                 |  174     46      0.146    0.039         0.294    0.077
  2                 |  172     57      0.145    0.048         0.291    0.095
  3                 |  104     66      0.088    0.056         0.176    0.111
  4                 |   92    101      0.077    0.085         0.156    0.169
  5                 |   41    154      0.035    0.130         0.069    0.258
  6                 |    8    173      0.007    0.146         0.014    0.290
  Total             |  591    597      0.497    0.503         1.000    1.000

  Response Category |  Cumulative Conditional
                    |      Probability
                    |     Noise       Signal
  ------------------+-----------------------
  1                 |     0.294       0.077
  2                 |     0.585       0.173
  3                 |     0.761       0.283
  4                 |     0.917       0.452
  5                 |     0.986       0.710
  6                 |     1.000       1.000
  Total             |

Initial Estimates of Parameters: Gaussian Model

                 Standard                            Standard
                 Deviation          Mean             Deviation
  Mean(Noise)    (Noise)     (Signal+Noise)    (Signal+Noise)
  -----------------------------------------------------------
     0.000        1.000          1.495              1.392

  D-Prime   D Sub-A   Sakitt D   SD-Ratio   ROC Area
  --------------------------------------------------
    1.495     1.234     1.267      1.392      0.808
```

Upper Category Boundaries

-0.523	0.204	0.706	1.366	2.229

Goodness of Fit

Log Likelihood	Chi-square (df:3)	p-value
-1921.419	2.364	0.500

Iterative Maximum-Likelihood Estimation of Parameters with Tolerance :0.001

Iteration	-Log Likelihood	D-Prime	SD-Ratio	Category Boundaries			
0	1921.419	1.495	1.392	-0.523	0.204	0.706	1.366
1	1920.995	1.508	1.409	-0.536	0.200	0.705	1.360
2	1920.986	1.522	1.417	-0.534	0.204	0.711	1.367
3	1920.985	1.518	1.416	-0.533	0.204	0.710	1.366
4	1920.985	1.518	1.416	-0.533	0.204	0.710	1.366
5	1920.985	1.519	1.417	-0.533	0.204	0.710	1.366
6	1920.985	1.519	1.417	-0.533	0.204	0.710	1.366

Iteration	
0	2.229
1	2.283
2	2.295
3	2.294
4	2.294
5	2.294
6	2.294

Final Parameter Estimates using Upper Category Boundaries: Gaussian Model

Mean(Noise)	Standard Deviation (Noise)	Mean (Signal+Noise)	Standard Deviation (Signal+Noise)
0.000	1.000	1.519	1.417

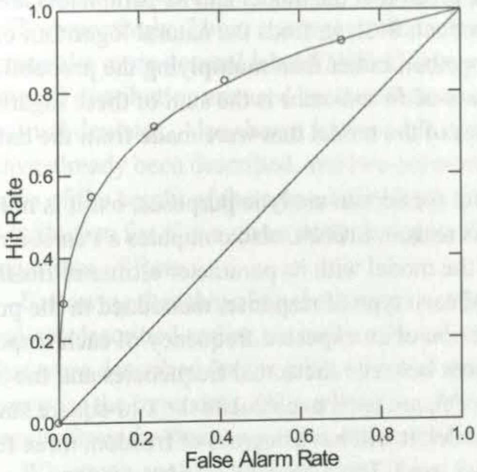
Category Label	FAR	HR	FINV(FAR)	FINV(HR)	Upper Boundary	Beta	Log(Beta)
1	0.706	0.923	-0.541	-1.425	-0.533	0.285	-1.255
2	0.415	0.827	0.216	-0.944	0.204	0.468	-0.758
3	0.239	0.717	0.711	-0.574	0.710	0.771	-0.260
4	0.083	0.548	1.386	-0.120	1.366	1.785	0.579
5	0.014	0.290	2.210	0.554	2.294	8.437	2.133

D-Prime	D Sub-A	Sakitt D	SD-Ratio	ROC Area
1.519	1.239	1.276	1.417	0.809

Goodness of Fit

Log Likelihood	Chi-square (df:3)	p-value
-1920.985	1.497	0.683

Receiver Operating Characteristic (ROC) Plot



The meanings of the first two sections of the output are self-evident. The first is a report on the data read, and the second is a tabulation of frequencies and probabilities (relative frequencies) compiled from the input data. The final column of the frequency table lists *Cumulative Conditional Probabilities*. The false-alarm rates (*FAR*) and hit-rates (*HR*) shown later in the output are computed by subtracting these cumulative probabilities from unity. This results in *FAR* and *HR* being associated with the upper category boundary of a labeled category. The report on the data read and the frequency table are a standard part of the output for every model that can be used in SIGNAL.

The next section of the output, labeled *Initial Estimates of Parameters*, contains the detection parameters discussed earlier as well as a line of numbers labeled *Upper Category Boundaries*. The latter are the standard normal deviates (*z* scores) that correspond to the area of the noise distribution that is above the upper boundary of each successive response category. Notice that there is one fewer of these than there are categories. This section of the output is referred to as *Initial Estimates* because they are the starting estimates that the iterative procedure uses to compute maximum likelihood estimates of these same parameters.

At the end of the output for initial estimates, there is a line labeled *Goodness of Fit*. The data on this line indicate how well the initial parameter estimates account for the empirical data in your data file. The estimated parameters and category boundaries for the two normal distributions (*N* and $S + N$) allow the estimation of the probability that

a given response will occur given either an N trial or an $S + N$ trial. We compute these probabilities for each response that occurred on each trial of the experiment. The product of all these probabilities gives us the probability (likelihood) of obtaining the data that we in fact obtained, given that the model and its parameters are correct. Instead of computing this product, SIGNAL finds the natural logarithm of each probability and adds them together, rather than multiplying the probabilities themselves. The first goodness-of-fit indicator is the sum of these logarithms for the input data, given the estimates of the model that were made from the data. Thus, it is labeled *Log Likelihood.*

The log-likelihood is useful for certain analytic purposes, but it is not very intuitively appealing. For this reason, SIGNAL also computes a Pearson chi-square statistic indicating how well the model with its parameter estimates fits the input data. The theoretical probability of each type of response, mentioned in the preceding paragraph, allows the calculation of an expected frequency of each response for both N and $S + N$ trials. Differences between the actual frequencies and the expected frequencies, based on the model, are used to calculate the Chi-square statistic in the usual way. For the normal model, it will have degrees of freedom three fewer than the number of response categories used. This *Chi-square* value, its *Degrees of Freedom*, and *p-value* are shown along with the *Log Likelihood* as Goodness of Fit statistics. The *p-value* will be unity if the fit is perfect and will approach 0 for very bad fits to the data.

The next section of the output is a history of the iterative estimation of the model parameters. The value of *–Log Likelihood* is shown for each iteration along with the estimated values of the model parameters for the iteration. As you can see from the output, the value of *–Log Likelihood* decreases at each iteration until it levels off, and the program ceases to iterate. As the value of *–Log Likelihood* decreases, the likelihood of having gotten the data that were obtained increases, hence the term *Maximum-Likelihood Estimation*. When the program can no longer produce significant increases in this likelihood by adjusting the parameters and is not producing parameter values that differ much from iteration to iteration, it ceases. The letter D that appears in this numerical output should be interpreted in the same way as an E is when using scientific notation. Using D rather than E merely signifies that double-precision arithmetic is being used in the calculations.

As you can see in the output for the iterative estimations, SIGNAL estimates *D-Prime*, the *SD-Ratio* and the *Upper Category Boundaries* on each iteration. *D-Prime* could just as easily be labeled *Mean* $S + N$, and *SD-Ratio* could be labeled *Standard Deviation of* $S + N$ because we have assumed the mean and standard deviation of N to be 0 and 1, respectively. As stated earlier, the numbers for the upper category boundaries are standard normal deviates (z scores) relative to the N distribution.

Following the history of iteration is a table showing the final estimates of the parameters along with some other information.

In this table of final parameter estimates, you will see a column labeled *FINV(FAR)* and another labeled *FINV(HR)*. These are the z scores corresponding to the *FAR* and *HR*, respectively. These z scores are the inverse function of the *FAR* and *HR* values, hence the more general label *FINV*. This is very useful when models other than the normal distribution are used because then we are not necessarily dealing with standard normal deviates. Also shown in this table are the *Upper Category Boundaries*, which have already been described, and two columns labeled *Beta* and *Log(Beta)*. **Beta** is the ratio of the height of the normal distribution for $S+N$ to the height of the normal distribution for N at a given upper category boundary. **Log(Beta)** is the natural logarithm of Beta.

Following the table of final estimates are the computed values for all of the detection indices described earlier, as well as the values of the same goodness-of-fit measures that were described for the table of initial estimates. The plot shows the usual ROC curve for the input data. *HR* is plotted against *FAR*, and the theoretical ROC curve that results from the final parameter estimates is shown.

When you analyze data that have fewer than four response categories using the NORMAL model, you will notice that no iterations occur. This is because the *HR* and *FAR* data can always be fit perfectly by an algebraic procedure for fewer than four categories. There are not enough degrees of freedom in the data to allow any error of estimation. Thus, for these cases, all that you will get is a table of final estimates, and the goodness-of-fit measures will show a perfect fit.

Example 2
Nonparametric Model for Signal Detection

If you use the NPAR option of the ESTIMATE statement, you will get some very simple output. In addition to the data report and the frequency table, you will get the *HR* and *FAR* for each category and the area under the nonparametric ROC function. This ROC is constructed by connecting the empirical points on the graph with straight lines. The area referred to is the area to the right and below the function defined by these lines. Bamber (1975) showed that this nonparametric ROC was essentially the same thing that mathematicians call an *ordinal dominance graph*. He finds that the area under such a graph is closely related to the Mann-Whitney U statistic, thus enabling hypotheses about such an area to be tested.

The nonparametric model is a simple way to get a quick look at your data, and if you believe that the assumptions of any of the parametric models that you can use are not justifiable for your data, then NPAR is the model for you.

The input is:

```
USE SWETSDTA
SELECT SUBJ=1
SIGNAL
MODEL RATING=SIGNAL
FREQ COUNT
ESTIMATE / NPAR
```

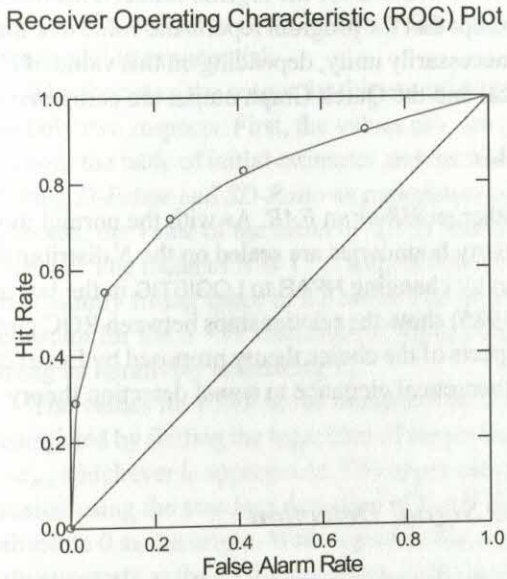
The output is:

```
Data for the following results were selected according to
      SELECT SUBJ=1

Case frequencies determined by value of variable COUNT

  Number of Stimulus Events (Cases) Responded to      : 1188.000
  Number of Detectors (variables) Observing an Event :    1.000
  Number of Response Categories Used                  :    6.000
  Number of Responses to Noise Events                 :  591.000
  Number of Responses to Signal Events                :  597.000
  Total Number of Responses                           : 1188.000
  Number of Instances of Missing Data                 :    0.000
```

Response Category	Frequency Noise	Frequency Signal	Joint Probability Noise	Joint Probability Signal	Conditional Probability Noise	Conditional Probability Signal
1	174	46	0.146	0.039	0.294	0.077
2	172	57	0.145	0.048	0.291	0.095
3	104	66	0.088	0.056	0.176	0.111
4	92	101	0.077	0.085	0.156	0.169
5	41	154	0.035	0.130	0.069	0.258
6	8	173	0.007	0.146	0.014	0.290
Total	591	597	0.497	0.503	1.000	1.000

Response Category	Cumulative Conditional Probability Noise	Cumulative Conditional Probability Signal
1	0.294	0.077
2	0.585	0.173
3	0.761	0.283
4	0.917	0.452
5	0.986	0.710
6	1.000	1.000
Total		

```
Nonparametric Analysis using Upper Category Boundaries

  False-alarm Rates for Successive Categories

  0.706   0.415   0.239   0.083   0.014

  Hit Rates for Successive Categories

  0.923   0.827   0.717   0.548   0.290

Area under ROC  :0.803
```



Example 3
Logistic Model for Signal Detection

This model uses the logistic distribution as the model for the N and $S+N$ distributions. The cumulative probability function for the logistic distribution is $1/(1+exp(-Y))$, where Y is the random variable on which the distribution is defined. In SIGNAL, Y is replaced by $c \times X$, where c is a scaling constant, and X is the decision axis of the detection model. The default value of c is $\pi/(\sqrt{3})$, which is approximately 1.81. This has the effect of making the variance of X equal to one. Thus, with the default in effect, the standard deviation of the N distribution will be 1.00. You can set the value of c to be any positive number because c is an option that can be appended to the ESTIMATE command for this model.

You might very well want to use 1.7 for c because, if you do, the cumulative probabilities for the N distribution will differ from standard normal (Gaussian) probabilities by less than 0.01 for all values of X (with a mean of 0). The standard deviation of X for the N distribution will not be unity, however. It will be equal to $\pi/(1.7\sqrt{3})$, or about 1.07. Thus, the logistic distribution is a good approximation for the normal, and it is mathematically more tractable.

The format of the output from SIGNAL for the logistic model does not differ from that for the normal model, except that the program reports the value of c that was used, and the variance of N is not necessarily unity, depending on this value of c. The values of *FINV* in both the numerical and the Quick Graph output are computed as:

$$FINV = \ln((1-p)/p)/c$$

where the probability p is either an *HR* or an *FAR*. As with the normal model, the values shown for upper category boundaries are scaled on the N distribution. You can examine the input and output by changing NPAR to LOGISTIC in the last example.

Treisman and Faulkner (1985) show the relationships between ROC curves derived from a logistic model and aspects of the choice theory proposed by Luce (2005). Thus, the use of this model lends theoretical elegance to signal detection theory.

Example 4
Negative Exponential Model for Signal Detection

The negative exponential density function is algebraically identical to a chi-square density function with two degrees of freedom. However, its simple properties and usefulness justify treating it as a separate model. A simple (somewhat limited) way to think of the distribution is to imagine a detector that on each trial of an experiment receives two random observations from a normal distribution that is either N or $S + N$ in nature. The detector is built to compute the variance, or sum of squared deviations of the two observations, and bases its response on that computation. Thus, it tries to distinguish $S + N$ from N, based on the variance of each. A chi-square distribution with two degrees of freedom is an appropriate model for such a detector.

The cumulative probability function for the negative exponential distribution is $1 - e^{cX}$, where X is a random variable (the decision axis in this case) and c is a constant. The mean and standard deviation of the distribution are both equal to $1/c$. Therefore, if we have an N distribution with the value of $c = c_N$ and an $S + N$ distribution with the value of $c = c_S$, D-Prime is $(c_N/c_S) - 1$ and the ratio of the $S + N$ standard deviation to the N standard deviation is c_N/c_S. A little algebra shows that the ROC is given by a power law: The *HR* is the *FAR* raised to the c_S/c_N power. The area under the ROC curve is $1/(1 + (c_S/c_N))$.

The implementation of this model in SIGNAL allows the user to choose a value of c for the N distribution. The program then finds the best-fitting value of c for the $S + N$ distribution given the input data. Thus, the user-supplied value of c is simply a scaling constant for the decision axis, and there is only one parameter left for the program to

estimate from the data (in addition to the category boundaries). The default value of c for the N distribution is unity. This is the only option for the MODEL statement when the model is exponential.

The format of the output for this model differs from the format of the normal model in only two respects. First, the values of c are given for the N and $S+N$ distributions in both the table of initial estimates and the table of final estimates. Second, rather than listing *D-Prime* and *SD-Ratio* as parameters being estimated during the iterative process, the value of the mean of $S+N$ and the value of the mean of N are listed instead. The mean of N is $1/c$ and, of course, remains constant during the iterations. It is simply filling space in the table. The mean of $S+N$ is $1/c_S$, where c_S is the constant for the $S+N$ distribution. Thus, iteratively estimating the mean is the same thing as iteratively estimating c_S.

The values for *FINV* in the numerical as well as the Quick Graph output are computed by finding the logarithm of the probability involved and dividing it by $-c_S$ or $-c_N$, whichever is appropriate. The upper category boundaries shown in the output are scaled using the standard deviation of the N distribution as the unit of measure and absolute 0 as the origin. With regard to the origin, notice that the linear ROC line always starts at the (0,0) coordinate of the plot, as it must for the exponential model.

As you will notice in the output, the degrees of freedom for the chi-square goodness of fit are equal to the number of response categories minus 2, rather than minus 3 as for the normal and logistic distributions. This is because there is one less parameter to estimate for the exponential model than for the other two.

The input is:

```
USE SWETSDTA
SELECT SUBJ=1
SIGNAL
MODEL RATING=SIGNAL
FREQ COUNT
ESTIMATE / EXPONENTIAL
```

The output is :

```
Data for the following results were selected according to
    SELECT SUBJ=1

Case frequencies determined by value of variable COUNT

  Number of Stimulus Events (Cases) Responded to     : 1188.000
  Number of Detectors (variables) Observing an Event :    1.000
  Number of Response Categories Used                 :    6.000
  Number of Responses to Noise Events                :  591.000
  Number of Responses to Signal Events               :  597.000
  Total Number of Responses                          : 1188.000
  Number of Instances of Missing Data                :    0.000
```

Response Category	Frequency Noise	Frequency Signal	Joint Probability Noise	Joint Probability Signal	Conditional Probability Noise	Conditional Probability Signal
1	174	46	0.146	0.039	0.294	0.077
2	172	57	0.145	0.048	0.291	0.095
3	104	66	0.088	0.056	0.176	0.111
4	92	101	0.077	0.085	0.156	0.169
5	41	154	0.035	0.130	0.069	0.258
6	8	173	0.007	0.146	0.014	0.290
Total	591	597	0.497	0.503	1.000	1.000

Response Category	Cumulative Conditional Probability Noise	Cumulative Conditional Probability Signal
1	0.294	0.077
2	0.585	0.173
3	0.761	0.283
4	0.917	0.452
5	0.986	0.710
6	1.000	1.000
Total		

Initial Estimates of Parameters: Exponential Model

Multiplicative Constant for Noise :1.000
Multiplicative Constant for Signal+Noise :0.258

Mean (Noise)	Standard Deviation (Noise)	Mean (Signal+Noise)	Standard Deviation (Signal+Noise)
1.000	1.000	3.870	3.870

D-Prime	D Sub-A	Sakitt D	SD-Ratio	ROC Area
2.870	1.015	1.459	3.870	0.795

Upper Category Boundaries

0.346 0.871 1.424 2.480 4.333

Goodness of Fit

Log Likelihood	Chi-square (df:4)	p-value
-1927.548	15.180	0.004

Iterative Maximum-Likelihood Estimation of Parameters with Tolerance :0.001

Iteration	-Log Likelihood	Mean (Signal+Noise)	Mean (Noise)	Category Boundaries			
0	1927.548	3.870	1.000	0.346	0.871	1.424	2.480
1	1922.809	3.998	1.000	0.342	0.846	1.407	2.476
2	1922.807	4.008	1.000	0.343	0.847	1.407	2.477
3	1922.807	4.008	1.000	0.343	0.847	1.407	2.477
4	1922.807	4.009	1.000	0.343	0.847	1.408	2.476

Iteration	
0	4.333
1	4.856
2	4.869
3	4.869
4	4.871

```
Final Parameter Estimates using Upper Category Boundaries: Exponential Model

Multiplicative Constant for Noise          :1.000
Multiplicative Constant for Signal+Noise   :0.249
```

Mean (Noise)	Standard Deviation (Noise)	Mean (Signal+Noise)	Standard Deviation (Signal+Noise)
1.000	1.000	4.009	4.009

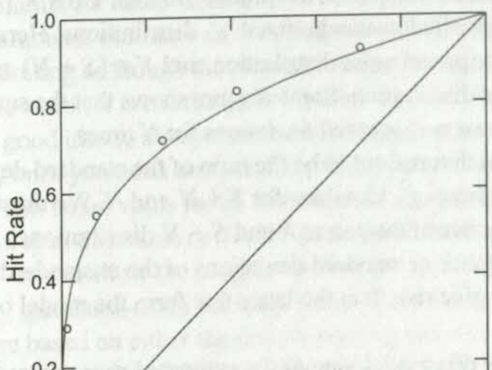
Category Label	FAR	HR	FINV(FAR)	FINV(HR)	Upper Boundary	Beta	Log(Beta)
1	0.706	0.923	0.349	0.080	0.343	0.323	-1.131
2	0.415	0.827	0.881	0.189	0.847	0.471	-0.753
3	0.239	0.717	1.433	0.333	1.408	0.718	-0.332
4	0.083	0.548	2.490	0.602	2.476	1.600	0.470
5	0.014	0.290	4.302	1.239	4.871	9.651	2.267

D-Prime	D Sub-A	Sakitt D	SD-Ratio	ROC Area
3.009	1.030	1.503	4.009	0.800

Goodness of Fit

Log Likelihood	Chi-square (df:4)	p-value
-1922.807	5.579	0.233

Receiver Operating Characteristic (ROC) Plot



Egan (1975) discusses the negative exponential model at some length. He points out its relationship to the Rayleigh distribution and notes that it represents the probability distribution of a randomly selected sinusoid in the Rice model of Gaussian noise.

Example 5
Chi-Square Model for Signal Detection

We can think of the chi-square model as a generalization of the exponential model that was just discussed. Imagine that the hypothetical detector now receives k random observations of N, or k random observations of $S+N$ on a trial, and that N and $S+N$ are normally distributed. As before, the detector bases its response on the sums of squared deviations for the k observations. The appropriate model for such a detector is the (unstandardized) chi-square distribution with k degrees of freedom (or $k-1$ degrees of freedom if the detector does not know and must estimate the true means of N and $S+N$ from the data). For SIGNAL, we assume that k is the same for both $S+N$ and N trials.

Let us designate a χ^2 variable that represents the sum of squared deviations for k observations from the N distribution divided by the population variance of the parent (normal) distribution as *CSN*. Let *CSS* be the corresponding χ^2 variable for the $S+N$ trials. The distributions of sums of squared deviations are then $CSN \times Var(N)$ and $CSS \times Var(S+N)$, the so-called unstandardized χ^2 distributions. Here, $Var(N)$ represents the variance of the parent noise distribution, and $Var(S+N)$ represents the variance of the signal+noise distribution. Some algebra shows that the sum of squared deviations for $S+N$ is the sum of squared deviations for N times $Var(S+N)/Var(N)$, which turns out to be the ratio of the standard deviations (not variances) of the unstandardized χ^2 variables for $S+N$ and N. We must be careful here to distinguish the variances of the parent N and $S+N$ distributions, $Var(N)$ and $Var(S+N)$, from the variance or standard deviations of the unstandardized χ^2 distributions to which they give rise. It is the latter that form the model of what the detector is doing.

In the SIGNAL output for this model, one of the estimated parameters is called *SD-Ratio*. This is the ratio of the standard deviation of the $S+N$ unstandardized χ^2 variable to the standard deviation of the N unstandardized χ^2 variable. As stated above, this is also the ratio of $Var(S+N)$ to $Var(N)$. The unstandardized χ^2 for $S+N$ is a constant times the unstandardized χ^2 for N. The constant is the *SD-Ratio*. The means and standard deviations that SIGNAL prints for the N and $S+N$ distributions are based on the assumption that $Var(N)$ is unity. The mean for the N

unstandardized χ^2 is then simply the degrees of freedom, *k*. The mean for the $S + N$ unstandardized χ^2 is then *SD-Ratio* times *k*. Of course, the *SD-Ratio* then has unity as its denominator so that it is just equal to the standard deviation of the unstandardized χ^2 distribution for $S + N$.

The other parameter that can be estimated by the program is the correct degrees of freedom *(df)* to use. It is allowed to be a non-integer value. You do not have to allow SIGNAL to try to estimate *df*. If you want to fix *df* at some value, you can use an option of the MODEL command to do this. For example, if you want to fix *df* at 4, you should type:

```
ESTIMATE / CHISQUARE,DF=4
```

Consequently, the *df* will not change during the iterations. You can fix *df* at any positive value. There is another option available here that will give you some flexibility about the *df*. For example, you may type:

```
ESTIMATE / CHISQUARE,START,DF=12.8
```

This will cause SIGNAL to use 12.8 (or any other positive number that you type) as a starting value for the iterations. SIGNAL will (probably) move away from this starting value during the iterations in an attempt to find the best-fitting value for *df*.

There is a potential problem with allowing SIGNAL to do this. The procedure for trying to iteratively determine *df* seems particularly prone to the problem of local minima for some kinds of data. This means that, at some point in the process, the parameter estimates do not change much for a few iterations, and so the program stops iterating as though the minimum value for the *–Log Likelihood* has been found. However, it is not the global minimum but a local minimum that has been found. It is a good idea to always use the START option to start the program at a different value of *df* once you think that you have found the maximum likelihood solution. There may be a still better value for *df*. Better yet, if you have some idea of what *df* should be, either fix *df* at this value or START at it. The default starting value for *df* is 10.

If you allow the program to find the *df*, a degree of freedom is lost for the Pearson χ^2 goodness-of-fit statistic. This will not occur for the initial estimates because these are based on either the default starting value or a number that you assign to *df*. However, if you let the program iteratively find *df*, then the Pearson χ^2 will show one less degree of freedom for the final estimates.

These degrees of freedom are the number of categories minus 3 if the program estimates *df*, or the number of categories minus 2 if you fix *df* at a value. There is an unresolved theoretical problem here. If you use three categories of response and allow SIGNAL to iteratively find *df*, there will be one degree of freedom for the initial Pearson statistic (the starting value for *df* is given) and zero degrees of freedom for the final

Pearson statistic. Zero degrees of freedom would seem to imply that a perfect fit is a necessity, but it is not, as can easily be demonstrated with an example. In this case, the program still computes the empirical value for the Pearson statistic and finds the probability of that value based on one degree of freedom. However, the printout will show that there are zero degrees of freedom. There seems to be no resolution for this problem at present. This model also is a bit slower in execution than some others because of the complexity of finding inverse values for χ^2 probabilities, and because of the necessity to use an iterative technique to measure the area under the ROC.

As with the other models, the values for the upper category boundaries are measured on the *N* distribution. The printout for this model is so similar to the others that have been described that it will not be discussed further here. For further information on this model and ways in which it can be used, see Egan (1975).

Example 6
Poisson Model for Signal Detection

This model is appropriate when the detector is basing a response on a count of rare events of some kind that occur during a trial. On *N* trials, only a very few of the events are liable to occur; and on $S + N$ trials, more of the events occur, although they are still rare. Think, for example, of a rare form of bacteria that is present to some small degree in every person. Suppose that the presence of a certain disease is indicated by a small increase in the count of this bacteria as seen on a microscope slide. A slide containing a very small number of these organisms is considered to be from a normal person (a noise trial), and a slide with more of the bacteria is considered to have come from a diseased person (the signal+noise condition). It must be decided on the basis of small differences in count whether a person is diseased or not (or a rating scale of the likelihood of disease could be used). When the number of possibilities for bacteria is large, as on the slide, but the probability of finding very many is small, the Poisson model is appropriate.

The Poisson is a discrete distribution with probabilities defined only for the non-negative integers. This would seem to lead to a theoretical ROC function that was composed of discrete points on the graph, one for each integer. However, Egan (1975) and others have argued that a guessing strategy when an ambiguous count is received by the detector allows us to "close" the ROC by connecting the points with straight lines.

If the detector were mechanical or electrical, you would have to assume that when an ambiguous count was received on a trial, the detector would sometimes act as though the next highest integer was appropriate and sometimes act as though the next

lowest was appropriate, perhaps with unequal probabilities. This behavior allows us to close the ROC with the aforementioned straight lines. This is the approach taken in SIGNAL.

The decision axis is the scale of non-negative real numbers; and while the probabilities of various counts theoretically can occur only at integer values, the closing of the ROC implies that boundaries between response categories can occur at any real number. Thus, the scale of the decision axis is fixed by the non-arbitrariness of the counting numbers. The question then becomes, "What two Poisson distributions defined on these numbers best fit the response data given?"

Formulas for the Poisson distribution are given in many statistics texts. The mean (λ) of the Poisson distribution is the average number of occurrences per trial of the event being counted. A trial can be a spatial and/or temporal entity. The variance of the Poisson is also λ. Thus, there are two model parameters to deal with here (in addition to category boundaries): the mean of the N distribution and the mean of the $S + N$ distribution.

If you fix one of these means at some value, a priori, then there is only the other mean and the category boundaries for SIGNAL to estimate. That is exactly what one of the ESTIMATE options allows you to do. You can specify that the mean of the N distribution be fixed at some value. For example, if you type:

```
ESTIMATE / POISSON,MEAN=4
```

the program will fix the mean of the N distribution at 4 and then estimate the value for the $S + N$ mean that gives the best fit to your data. The two means completely specify the two distributions. The default value for the MEAN option is 5.

The START option used here works in a manner similar to that described for the chi-square model. It makes the initial value of the mean of N equal to the set value. The program would then include values of the mean of N in the iterative process, trying to find a best-fitting value for the mean (that is, trying to find the most appropriate Poisson distribution). The same comments that were made above in the chi-square model regarding the Pearson fit statistic apply here as well.

Like the chi-square model, the iterative routine seems to be susceptible to a local minimum problem for many Poisson data sets. Thus, the best strategy, if you do not know the value of the mean for N, is to try a wide range of fixed values for it in successive runs of the model. Then pick the fixed value that gave you the lowest value of *–Log Likelihood* and use it along with the START option to allow the program to iterate near that value.

By the time you have read this far, the nature of the output should be self-evident to you. It is very much the same for all of the models discussed. The program is somewhat

slower for the Poisson because of the iterative techniques that are needed to be used to find *FINV* for *HR* and *FAR* and to find the area under the ROC.

Example 7
Gamma Model for Signal Detection

Suppose that in an experiment, the N and $S+N$ trials can be described as a Poisson process as described for the Poisson model. That is, a small number of discrete, countable, and rare events occur on each trial. But now suppose that the detector adopts or is programmed for the following strategy: The detector uses the time required to accumulate a fixed number of the rare events as the basis of the response. If that fixed number accumulates very slowly, the detector gives a "noise-like" response. If the fixed number accumulates more rapidly, the detector gives a "signal-like" response. Thus, the detector's response (binary or rating category) is based on time—the time it takes to accumulate a predetermined number of discrete events. Notice that the length of a trial is determined by the detector, as opposed to the Poisson-counting observer described above, who counts events during a fixed-interval trial. In the former case, the gamma distribution is an appropriate detection model.

Formulas for the gamma distribution are found in advanced statistics textbooks. Suffice it to say here that the distribution has two parameters: m and r. Let us call the number of events that the detector is waiting to accumulate r. The mean of the gamma distribution is r/m so that m times the mean is r, and m is then a scaling constant. For a detection problem, there is only one value of r; but there are two values of m: one for the N distribution and one for the $S+N$ distribution. Let us call these m_o and m_1, respectively. The mean of the time that it takes for r (Poisson) events to accumulate if the N process is in effect is then r/m_o, and the mean of the time that it takes for r events to accumulate if the $S+N$ process is in effect is r/m_1. The variance of a gamma distribution is r divided by m^2, so knowing r and m defines both the mean and variance.

Thus, in addition to category boundaries, there are three parameters for our model: R, M0, and MI. In SIGNAL, we fix M0 at a value predetermined by the program (the default value) or by the user via the option described below. The default value is unity. If you want to change the value of M0, for example, to 3, then type:

```
ESTIMATE / GAMMA,M0=3
```

The value of MI is then estimated by the program after M0 is determined. Note that M0 is never estimated. It either keeps its default value or the value that you assign. These

values, along with R, determine the means and variances of the *N* and $S + N$ distributions. The value of R has a default value of 5. You can change it in a manner similar to that for M0. If you want to change both M0 and R, you must list them both in the same MODEL statement.

If you do not exercise the option to fix R, then the default starting value will be used and the program will estimate a value of R at every iteration. You can also choose to let this happen but pick your own starting value for R, just as in the previous two models. For example, you could type:

```
ESTIMATE / GAMMA,M0=3,START
```

to accept the default value of R and let M0 change over iterations. All of the discussion of local minima in the previous two models applies here as well. Do not let the program do your thinking for you.

If you think about it, you will realize that the *N* and $S + N$ distributions have to be reversed from their usual positions on the decision axis for this model. The *N* trials result in longer waiting times, and the $S + N$ trials result in shorter waiting times. Also, for the same reason, *HR* and *FAR* are in the lower tail instead of the upper tails of the distributions in this case. This has all been taken care of for you with SIGNAL. Everything that needs to be reversed has been reversed so that, for example, you do not get negative values of *D-Prime* when you should not. (That doesn't mean that negative values cannot occur.)

Again, the output is in the same approximate format as for the other models already described. The few differences should be self-evident. It should also be evident that this model was specifically designed to handle waiting-time data. If you use it as a general GAMMA model, you will have to remember all of the design features mentioned here, especially the reversal of the direction of the decision axis and the fact that *HR* and *FAR* are computed from the lower tails of the distributions.

For more in-depth discussion of this model, you should again consult Egan (1975), who also makes some interesting comparisons of Poisson counting detectors versus gamma timing detectors.

Computation

Algorithms

The algorithm used to minimize negative log-likelihood is an adaptation of the Nelder-Mead simplex method as presented by O'Neill (1985). This method does not require derivatives, making it useful for all of the present models simultaneously. It is, however, less time-efficient than methods that use derivatives.

The area under the ROC is not directly computable for certain of the models (CHISQ, POISSON, and GAMMA). An algorithm was written to approximate this area by successively dividing it into smaller and smaller trapezoids and using the trapezoidal rule to accumulate the area. On successive iterations, the area is subdivided into trapezoidal panels, first two, then four, then eight, etc. If the increase in accumulated area is less than the value of CONVERGE from one iteration to the next, the subroutine ceases and returns the most recent value of the area. If, after 512 panels have been constructed, the stopping rule has not been met, the routine prints a warning message, returns the most recent area estimate, and ceases.

The initial estimates for the NORMAL and LOGISTIC models are obtained by finding the eigenvector of the *FINV(HR)* and *FINV(FAR)* vectors. The category boundaries are located by projecting the data points onto this vector and then scaling to the units of the *N* distribution. Similar methods are used for the other models, with the restriction that the least-squares vector on which the boundaries are located must pass through the point 0,0 on the linear ROC.

Missing Data

Missing data are treated as though the trial or trials that are missing the data did not exist only for the particular detector missing the data.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology,* 12, 387–415.

Coombs, C. H., Dawes, R. M., and Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Green, D. M. and Swets, J. A. (1989). *Signal detection theory and psychophysics*. Reprint ed. Los Altos Hills, CA: Peninsula Publishing.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143, 29–36.

Kraemer, H. C. (1988). Assessment of 2 x 2 associations: Generalization of signal detection methodology. *The American Statistician,* 42, 37–49.

Luce, D. (2005). *Individual choice behavior*. Mineola, NY: Dover.

O'Neill, R. (1985). Function minimization using a simplex procedure. In Griffiths, P. and Hill, I. D. (eds.), *Applied statistics algorithms*. Chichester, England: Ellis Horwood Limited. 79–87.

Peterson, W. W., Birdsall, T. G., and Fox, W. C. (1954). The theory of signal detectability. *Institute of Radio Engineers Transactions,* PGIT-4, 171–212.

Sakitt, B. (1973). Indices of discriminability. *Nature,* 241, 133–134.

Simpson, A. J. and Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin,* 80, 481–488.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin,* 99, 110–117.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Swets, J. A, Tanner, W. P., and Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review,* 68, 301–340.

Treisman, M. and Faulkner, A. (1985). On the choice between choice theory and signal detection theory. *Quarterly Journal of Experimental Psychology,* 37A, 387–405.

Chapter
11

Smoothing

Leland Wilkinson

The SMOOTH module applies nonparametric smoothers for data exploration in two or three dimensions. In nonparametric smoothing, a weighted function of a data subset provides a local estimate for a region. Because each region receives a smooth estimate, the agglomeration of these estimates captures local variations well without the need for complicated models or additional parameters found in parametric smoothing.

Constructing a nonparametric smoother involves:

- Specifying the size of the estimation regions. Estimation regions are defined by the number of neighboring points or as a fixed range of data.
- Defining the weighting function. Smoothing offers seven functions for data weighting: Epanechnikov, biweight, triweight, tricube, uniform, Gaussian, and Cauchy.
- Assigning a method of combining the weighted observations. Estimates can be computed as means, trimmed means, medians, polynomial regression estimates, or robust estimates.

The combinations of estimation window, weighting function, and smoothing method result in 126 possible nonparametric smoothers. For each smoother, you can estimate values at either specified gridpoints or at the predictor data values, saving the results to SYSTAT files for subsequent analyses.

Resampling procedures are available in this feature.

Statistical Background

Smoothers fit functions to data. Nonparametric smoothers fit functions to overlapping subsets of data. Each subset receives its own fit so that the overall smoother adapts to local variations in the data. Unlike parametric smoothers (such as linear or polynomial regression) nonparametric smoothers have no global model equation or simple parameters; instead, they have a collection of local estimates. Consequently, they are designed for data exploration and local prediction rather than parametric modeling. There is a large statistical literature in this area (e.g., Härdle, 1990; Hastie and Tibshirani, 1990; Green and Silverman, 1994; Fan and Gijbels, 1996; Simonoff, 1996). It is worth consulting one of these references before using SMOOTH.

Tukey (1977) used the word *smooth* for a procedure that describes the given data as follows:

data = smooth PLUS rough

This equation is a species of a more general Tukey paradigm:

data = fit PLUS residual

If we had perfect knowledge of the process that led to our observed data values, then we might construct a complete description of our data. We are usually safer in regarding our fit as an incomplete description, however, as Tukey says:

data = incomplete description PLUS residual

As we shall see, smoothing in practice inevitably produces an incomplete description. It involves a variety of trade-offs and requires careful judgment. This introduction will summarize several of these trade-offs.

The Three Ingredients of Nonparametric Smoothers

Nonparametric smoothers are usually assembled from three ingredients: 1) A **kernel function**, 2) a **bandwidth function**, and 3) a **smoothing function**. The kernel function is a probability function that is used to weight points in the computation of each local smoothing estimate. Points farther from the location of the estimate are usually weighted less than points nearer. The bandwidth function determines the size of the region over which each local estimate is computed. It does this by setting the spread or width of the kernel function. Finally, the smoothing function is the method for

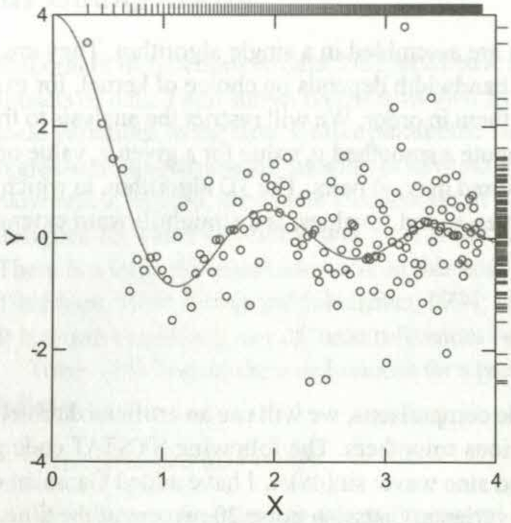
computing a smoothed estimate over the subset of points lying within the smoothing window.

These three ingredients are assembled in a single algorithm. They are interdependent (choice of bandwidth depends on choice of kernel, for example), although we will describe them in order. We will restrict the analysis to the case of 2D smoothing, where we compute a smoothed y_s value for a given x_s value on the basis of a collection of points measured in (x, y) pairs. The 3D algorithm, in which we compute a smoothed z_s value for given x_s and y_s values, is a straightforward extension of the same principles.

A Sample Data Set

In order to make systematic comparisons, we will use an artificial data set that exploits differences among the various smoothers. The following SYSTAT code generates a data set based on a damped sine wave: $\sin(4x)/x$. I have added Gaussian error to this function as well as higher-variance Gaussian noise 20 percent of the time. Also, I have applied the square-root transformation to the x values in order to make them unequally spaced. This will become relevant in comparing fixed-width windows with k nearest neighbor windows. The following plot shows 150 cases generated by this program, the function underlying the process, and the marginal distributions of the cases displayed as stripes at the borders.

```
NEW
RSEED 1357
REPEAT 150
LET X=SQR(CASE/10)
LET Y=SIN(4*X)/X+ZRN()/2+1.5*ZRN()*(URN()>0.8)
ESAVE DAMP1
USE DAMP1
BEGIN
PLOT Y*X/XMIN=0,XMAX=4,YMIN=-4,YMAX=4,BORDER=STRIPE
FPLOT Y=SIN(4*X)/X ; XMIN=0,XMAX=4,YMIN=-4,YMAX=4,
AX=0,SC=0,COLOR=RED
END
```

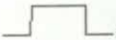


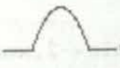
It is tempting to use this plot as a criterion for evaluating the goodness of the smoothing methods we will try, and to some extent this makes sense. Some combinations of the three smoothing parameters we will review do better than others in approximating the known smooth function. We must keep in mind, however, that one example cannot suffice for a global evaluation. First of all, the sinusoidal function we are using has some characteristics (periodic form, for example) more suitable for some methods than for others. Other functions would be more suited to other smoothing methods. Secondly, there are formal analytic results given in the references that should take precedence over eye-balling graphics when deciding on parameter choices. Third, there is room in this chapter to present only a limited subset of possible smoothing parameter combinations. With 7 kernel functions, 2 window types for determining bandwidth, and 9 smoothing methods, there are 126 different smoothers available in this package. This figure does not include the possible choices for bandwidth itself.

Kernels

The first ingredient of a nonparametric smoothing is a kernel function. Although many weighting functions have been proposed for smoothers at one time or another by applied researchers, the statistical literature has focused on a set of related kernel

functions that are suited to formal analysis of their properties. They are shaped as follows:

 *uniform*: $f(x) = a : (-h \le x \le h)$, else 0

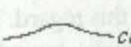
 *epanechnikov*: $f(x) = a(1-(x/h))^2 : (-h \le x \le h)$, else 0

 *biweight*: $f(x) = a(1-(x/h)^2)^2 : (-h \le x \le h)$, else 0

 *triweight*: $f(x) = a(1-(x/h)^2)^3 : (-h \le x \le h)$, else 0

 *tricube*: $f(x) = a(1-|x/h|^3)^3 : (-h \le x \le h)$, else 0

 *gaussian*: $f(x) = ae-^{(x/h)^2}$

 *cauchy*: $f(x) = a/(b+(x/h)^2)$

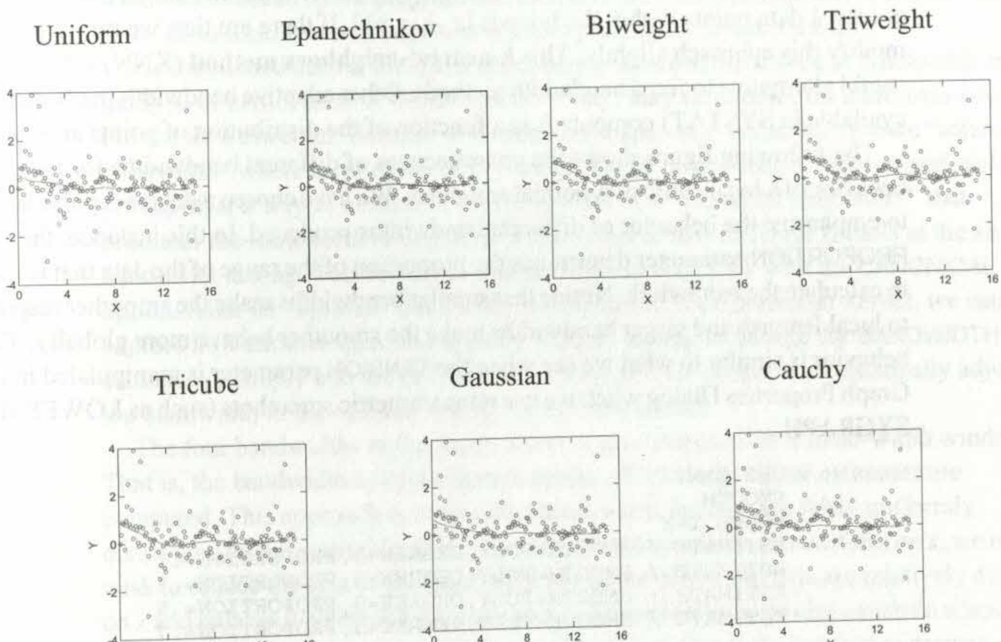
The constant *a* scales these formulas as probability kernel functions. This constant may be set to 1 for smoothers because it cancels out in the algorithms. The constant *h* is the bandwidth that determines the width of the window or, for kernels like *gaussian* and *cauchy* that have nonzero tails, the spread of the kernel. We can adapt these functions to 3D smoothing (assuming *x* and *y* are independent and identically distributed) by transforming them into polar coordinates. Non-circular 3D window functions are slightly more complex.

You may wish to try plotting these kernel functions in SYSTAT to learn more about the behavior of their parameters. The following commands are used to make the graphics on the left of each equation (from top to bottom):

```
fplot F=.5*(x>-1)*(x<1) ; xmin=-2,xmax=2,ymin=0,ymax=2
fplot F=.75*(1-x^2)*(x>-1)*(x<1);,
xmin=-2,xmax=2,ymin=0,ymax=2
fplot F=.9375*((1-x^2)^2)*(x>-1)*(x<1) ; xmin=-2,xmax=2,
ymin=0,ymax=2
fplot F=1.094*((1-x^2)^3) *(x>-1)*(x<1); xmin=-2,xmax=2,
ymin=0,ymax=2
fplot F=.864*((1-abs(x)^3)^3)*(x>-1)*(x<1);xmin=-2,xmax=2,
ymin=0,ymax=2
fplot F=.399*exp(-x^2) ; xmin=-2,xmax=2,ymin=0,ymax=2
FPLOT F=.32/(1+x^2) ; xmin=-2,xmax=2,ymin=0,ymax=2
```

The following figure shows the effect of these kernels on a running-mean smoother applied to our data. Some differences are subtle because of the similarity of shapes among some kernels, because of other default parameter settings such as bandwidth, and because of the particular data set we are using. Nevertheless, there are some differences worth noting. First, the Uniform kernel tends to downplay local variation because it weights all points in the window equally. By contrast, the Triweight weights points near the center of the window (where the estimate is computed) more heavily. Epanechnikov and biweight are between these two kernels in this regard. The Gaussian and Cauchy functions have infinitely long tails and central peaks, but when they are scaled to the intervals used for the other smoothers (as in the function plots shown above), they more closely resemble flatter weighting functions.

```
USE DAMP
SMOOTH
MODEL Y=X
ESTIMATE / SMOOTH=MEAN,KERNEL=UNIFORM
ESTIMATE / SMOOTH=MEAN,KERNEL=EPANECHNIKOV
ESTIMATE / SMOOTH=MEAN,KERNEL=BIWEIGHT
ESTIMATE / SMOOTH=MEAN,KERNEL=TRIWEIGHT
ESTIMATE / SMOOTH=MEAN,KERNEL=TRICUBE
ESTIMATE / SMOOTH=MEAN,KERNEL=GAUSSIAN
ESTIMATE / SMOOTH=MEAN,KERNEL=CAUCHY
```



Bandwidth

The second ingredient of a nonparametric smoothing is a bandwidth function. To compute a smoothed y_s value for a given x_s value, we need to compute the bandwidth function h at x_s so that we can scale the width of our kernel. There are several ways to do this.

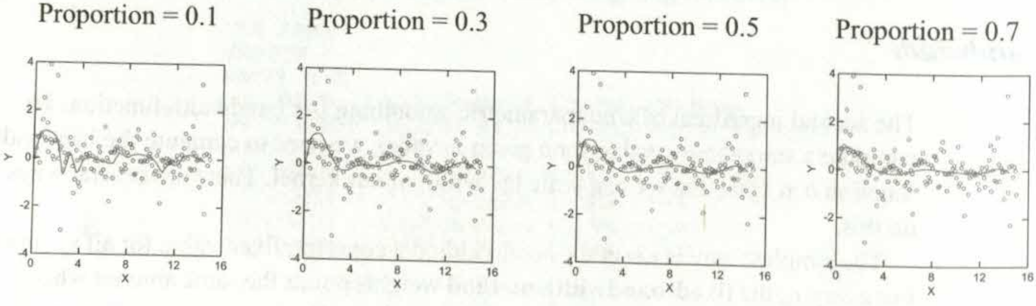
The simplest way is to set the bandwidth to a constant, fixed value for all x_s values. For every x_s, the **fixed-bandwidth method** weights points the same amount when they are the same distance from x_s. This method works well when the x coordinates of the data values are fairly uniformly distributed. It is the method underlying in popular time series smoothers such as moving averages.

Alternatively, we may devise an adaptive or variable method that yields a different bandwidth value for each x_s. One way to achieve this goal is to choose a subset size k less than or equal to the number of points n; this is the number of points on which we

want every smoothed value to be based. A popular choice is k=n/2. For any x_s, we then choose the k points that are nearest neighbors to x_s and then set the bandwidth to the distance between x_s and the farthest neighbor in this set. This assures us that there are at least k data points within the bounds [x_s-h, x_s+h]. If there are ties, we may have to modify this approach slightly. This **k nearest-neighbors method** (KNN) offers a useful alternative to fixed bandwidth methods. Other adaptive bandwidth methods (not available in SYSTAT) compute h as a function of the distribution of points around x_s.

The following figure shows the consequences of different bandwidths for the behavior of a local cubic polynomial smoother. We have chosen a degree 3 polynomial to emphasize the behavior of different bandwidths produced. In this instance, the PROPORTION parameter determines the proportion of the range of the data that is used to calculate the bandwidth. Notice that smaller bandwidths make the smoother respond to local features and larger bandwidths make the smoother behave more globally. This behavior is similar to what we see when the TENSION parameter is manipulated in the Graph Properties Dialog when we use nonparametric smoothers (such as LOWESS) in SYGRAPH.

```
USE DAMP
SMOOTH
MODEL Y=X
ESTIMATE / SMOOTH=POLY,DEGREE=3,PROPORTION=.1
ESTIMATE / SMOOTH=POLY,DEGREE=3,PROPORTION=.3
ESTIMATE / SMOOTH=POLY,DEGREE=3,PROPORTION=.5
ESTIMATE / SMOOTH=POLY,DEGREE=3,PROPORTION=.7
```



There is an extensive literature for determining "optimal" bandwidths for fixed window-width smoothers. This is summarized in the references. Erring in the direction of too-small a bandwidth introduces extraneous variance and erring in the direction of too-large a bandwidth favors bias in the smoother. It is often best to start with the default settings and to explore other nearby bandwidth values to see if systematic

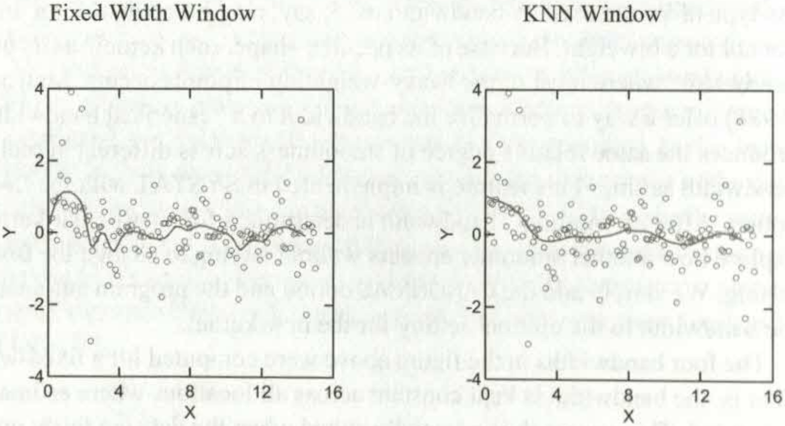
structure becomes apparent. The default settings in SYSTAT generally err on the side of large bias and small variance (over-smoothing). It is a good idea to observe the bandwidth chosen by the program and then to reduce it in subsequent runs to determine if the fit to local features improves without picking up random noise.

As mentioned above, there is a dependency between the choice of bandwidth and the type of kernel used. A bandwidth of .5, say, may be optimal for a uniform kernel, but not for a biweight. Because of its peculiar shape, each kernel has its own "effective bandwidth" where most of the heavy-weighting of points occurs. Marron and Nolan (1988) offer a way to normalize the bandwidth to a "canonical bandwidth" that produces the same relative degree of smoothness across different kernels at the same bandwidth setting. This feature is implemented in SYSTAT with the CANONICAL option. After an "optimal" bandwidth is determined for a particular kernel, we can explore how another smoother appears without having to change the BANDWIDTH setting. We simply add the CANONICAL option and the program automatically adjusts the bandwidth to the optimal setting for the new kernel.

The four bandwidths in the figure above were computed for a fixed-width window. That is, the bandwidth is kept constant across all locations where estimates are computed. This approach is especially suited when the data are fairly uniformly distributed on the x variable. When the data are substantially non-uniform on x, we may wish to reduce the size of the bandwidth in regions where the data are relatively dense on x and enlarge it where the data are sparse. A simple way to produce such an adaptive window width is to determine the bandwidth directly from the number of nearest neighbors. For example, we might decide that we want every estimate (point on the smoother) to be based on 30 points from our data set. In that case, the bandwidth will be wider for lower values of x (say, below 2) in our data set than for larger, because the points are concentrated at the higher values of x. The following figure shows the difference between the KNN (k nearest-neighbors) adaptive bandwidth and the fixed bandwidth.

The result is that the KNN window picks up more detail at the right-end of the smoother than the fixed window. Our choice of k (the number of neighbors) still determines whether we over- or under-smooth; the KNN window does not save us from having to make that decision. But it does provide more detail in areas of high data density. Combined with a robust smoothing method (such as Cleveland's LOESS), the KNN window can form the basis of an effective general-purpose smoother.

```
ESTIMATE / SMOOTH=POLY,WINDOW=FIXED,BANDWIDTH=.5
ESTIMATE / SMOOTH=POLY,WINDOW=KNN,NEIGH=30
```



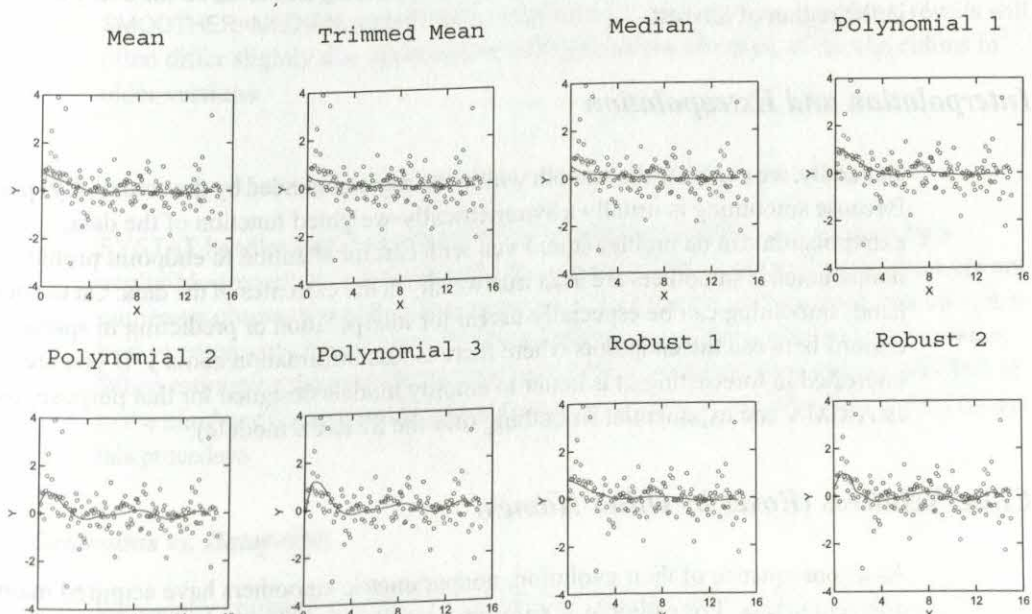
Smoothing Functions

The third ingredient in the smoothing algorithm is the smoothing function applied to data points using the kernel function to weight or select points. The smoothing function has only one requirement: it must yield a unique value y_s for any value x_s. In practice, the most popular functions are the same ones we use for statistical estimators: the mean, median, linear or polynomial regression estimate at x_s, and so on. In using these functions for smoothing, we must weight each point (x_i, y_i) by the kernel function value at $(x_s - x_i)$, but otherwise the computations are the same.

The literature suggests that polynomial smoothers tend to perform better, particularly at the endpoints of the smooth, than do mean smoothers. As the figure below shows, the MEAN smoothing function produces a smooth that tends to regress toward the mean of y at both ends. The TRIM and MEDIAN smoother are more resistant to outliers than the MEAN, but they have the same endpoint deficiency. The ROBUST polynomial smoothers have both good resistance to outliers and endpoint performance. The only drawback to using them is computation time. LOESS and other ROBUST smoothers require several additional passes through the data to achieve stable estimates.

The degree of the polynomial contributes to the local adaptiveness of the smoother. For curves that wiggle, degree 2 and 3 polynomials capture the bends better. It is best not to use higher polynomials if the data follow a monotonic path (increasing or decreasing), however. Higher degrees require more parameters and increase the risk of over-fitting. The following graphics show the different smoothers applied to our sample data.

```
SMOOTH
MODEL Y=X
ESTIMATE / SMOOTH=MEAN
ESTIMATE / SMOOTH=TRIM
ESTIMATE / SMOOTH=MEDIAN
ESTIMATE / SMOOTH=POLY,DEGREE=1
ESTIMATE / SMOOTH=POLY,DEGREE=2
ESTIMATE / SMOOTH=POLY,DEGREE=3
ESTIMATE / SMOOTH=ROBUST,DEGREE=1
ESTIMATE / SMOOTH=ROBUST,DEGREE=2
```



Smoothness

A remarkable aspect of the outcome of the nonparametric smoothing process is that we can produce a globally smooth curve from separate, local smooths. This happens most surely when we employ a bandwidth that weights a substantial subset of the data, and when we use a smoothing method that produces a continuous function (such as polynomial regression). Other combinations of kernels and smoothers produce a smooth global smooth for all practical purposes. This happens when kernel bandwidths are large enough to include sufficient data points for stable estimates and when smoothing functions are relatively continuous over the data.

Not all smoothing and kernel functions produce smooth smooths, however. If we use a smoothing function like the median or a relatively small bandwidth, we may find discontinuities in the global smooth. Discontinuities may not be undesirable. The point of smoothing is not to produce a smooth curve, but to produce an accurate summary y_s of the y_i data values whose paired x_i values lie near x_s, and to do so for every x_s value in the region of interest.

Interpolation and Extrapolation

Generally, we compute the smooth within the region bounded by the actual (x, y) pairs. Because smoothing is usually a symmetrically-weighted function of the data, extrapolation can be problematic. Even with careful attention to endpoint problems, nonparametric smoothers are least trustworthy at the extremes of the data. On the other hand, smoothing can be especially useful for interpolation or predicting in sparse x regions between the endpoints where there is less information about y. If you are interested in forecasting, it is better to employ models designed for that purpose, such as ARIMA and exponential smoothing (see the SERIES module).

Close Relatives (Roses by Other Names)

As a consequence of their evolution, nonparametric smoothers have acquired many different names. For example, a **running-means** or **moving-averages** smoother (Makridakis and Wheelwright, 1989) is a UNIFORM kernel smoother with MEAN smoothing function. The Nadaraya-Watson kernel smoother (Nadaraya, 1964; Watson, 1964) is an EPANECHNIKOV kernel with a MEAN smoothing function. Shepard's smoother (Shepard, 1965), sometimes called an **inverse-distance** smoother (McLain,

1974), is closely related to a CAUCHY kernel with MEAN smoothing function. The **distance-weighted least squares** (DWLS) smoother (McLain, 1974) is closely related to a quadratic polynomial smoother (SMOOTH=POLYNOMIAL, DEGREE=2) with a GAUSSIAN kernel. The **step** smoother (Cleveland, 1993) is a *k* nearest-neighbor smoother (WINDOW=KNN) with the number of nearest neighbors set to 1 (NEIGHBOR=1). The image-processing digital filter called a **discrete gaussian convolution** that is used to smooth black-and-white images (Gonzalez and Wintz, 1987) is a MEAN smoother with a GAUSSIAN kernel and KNN bandwidth function. It is a KNN method because the pixels on which it operates are evenly spaced, so it does weighted averages over a fixed number of pixels. Finally, Cleveland's LOESS smoother (Cleveland and Devlin, 1988) is a ROBUST smoother with DEGREE=1 or 2 and a TRICUBE kernel. The LOWESS smoother in SYGRAPH is based on an older, scatterplot smoothing version of this smoother. The LOESS option in SMOOTH is a shortcut to setting KERNEL=TRICUBE, SMOOTH=ROBUST, WINDOW=KNN. Finally, the running median smoother in Tukey (1977) is produced with WINDOW=KNN, SMOOTHER=MEDIAN, and KERNEL=UNIFORM. For all of these smoothers, results will often differ slightly due to parameter settings and peculiarities of the algorithms in older versions.

Ties

SYSTAT handles tied data in SMOOTH by randomly perturbing *x* values by a negligible amount before performing the smooth. This works well when there are not numerous observations at a given value of *x*. If your data contain many ties on *x*, it is best to preprocess them in the STATS module to aggregate *y* values at each *x* value. When estimating the smooth, use the mean of the *y* values at each *x* plus a WEIGHT set to the number of cases used to compute each mean. Hastie and Tibshirani (1990) discuss this procedure.

Gridpoints vs. Datapoints

We usually compute smoothed values on *y* at equally spaced values x_s that are not necessarily located at the data points. This yields a smoothed curve. The Quick Graph in SYSTAT plots the data and a curve through the smoothed points using a spline interpolator. It is often useful to SAVE the smoothed values so that they can be examined more closely and plotted directly in SYGRAPH. One of the examples below shows how to do this. Because they employ the additional step of spline-smoothing the

smoothed values (in order to save computation time), the 2D and 3D Quick Graphs are intended only for an initial glimpse at the smoothed results. You should always SAVE your results and plot them yourself after you are satisfied that a smooth is reasonable.

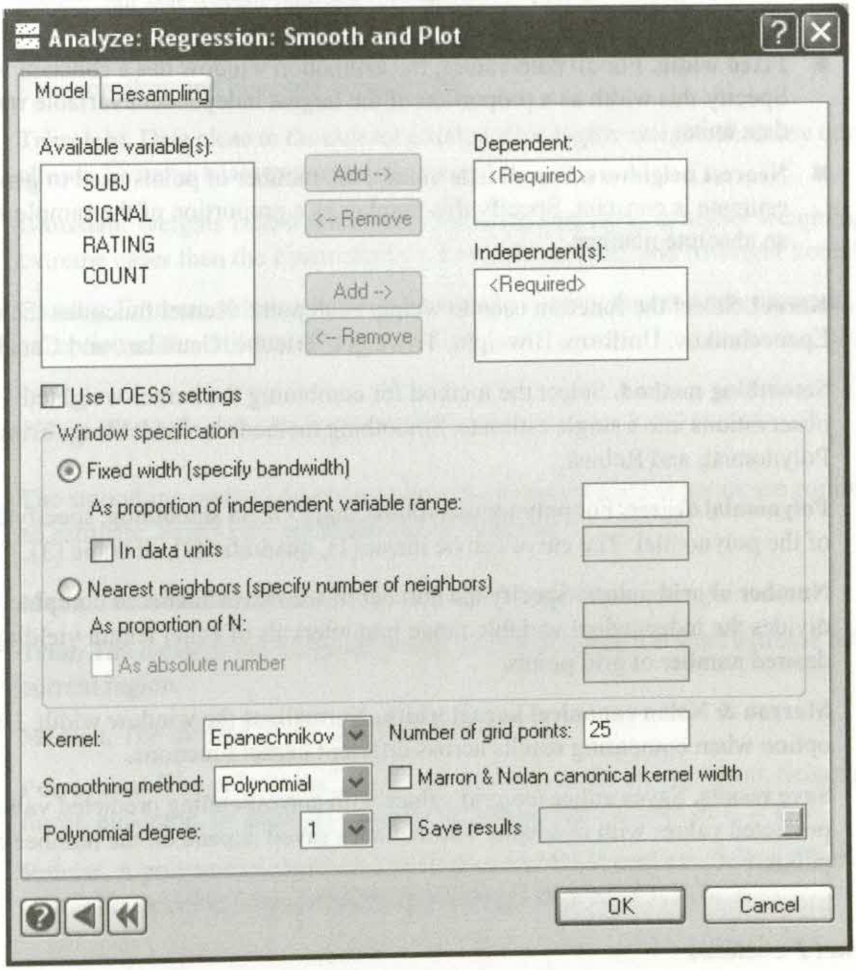
Often, we wish to plot residuals before settling on a particular smooth. This is done by setting GRID=0. In that case, the smoothed values are computed only at the values of the x_i data points. Examining residuals for specific smooths is as important in nonparametric as in parametric smoothing. You should look for the same features in the residuals in both cases: systematic departures from a rough horizontal band of residuals. A robust method such as LOESS will highlight outliers, however, while less resistant methods will mask them in the residuals. Of more concern than outliers is systematic trend or local variation that is not picked up by the smoother.

Smoothing in SYSTAT

Smooth & Plot Dialog Box

To fit a smoother, from the menus choose:

Analyze
  Regression
    Smooth & Plot...



Select the dependent and independent variables. Define the smoother by selecting from the following options:

Use LOESS settings. Uses the nearest 50% of the data for each estimate, employing a tricube kernel with robust smoothing. Unchecking this option allows you to specify a custom smoothing technique by selecting an estimation window, a kernel function, and a smoothing method.

Window specification. Define the region, or "window", containing the data to be used in calculating the smooth estimate for each point. Select either:

- **Fixed width.** For all data values, the estimation window has a constant width. Specify this width as a proportion of the largest independent variable range or in data units.
- **Nearest neighbors.** For all data values, the number of points used to generate the estimate is constant. Specify this number as a proportion of the sample size or as an absolute number.

Kernel. Select the function used to weight each point. Kernel functions include: Epanechnikov, Uniform, Biweight, Triweight, Tricube, Gaussian, and Cauchy.

Smoothing method. Select the method for combining the kernel-weighted observations into a single estimate. Smoothing methods include: Mean, Trim, Median, Polynomial, and Robust.

Polynomial degree. For polynomial, robust, and LOESS smoothing, specify the degree of the polynomial. The curve can be linear (1), quadratic (2), or cubic (3).

Number of grid points. Specify the number of smooth estimates to compute. SYSTAT divides the independent variable range into intervals of equal length yielding the desired number of grid points.

Marron & Nolan canonical kernel width. Normalizes the window width. Use this option when comparing results across different kernel functions.

Save results. Saves either the grid values with corresponding predicted values or the predicted values with residuals. The statistics saved depend on the number of grid points.

Kernel Functions

The kernel function determines the weights assigned to points used in calculating a particular smoothing estimate. Select one of the following functions:

Uniform. All data receive equal weights.

Epanechnikov. Data near the current point receive higher weights than what extreme data receive. This function weights extreme points more than the triweight, biweight, and tricube kernels, but less than the Gaussian and Cauchy.

Biweight. Data far from the current point receive more weight than the triweight kernel allows, but less weight than the Epanechnikov kernel permits.

Tricube. Data close to the current point receive higher weights than both the Epanechnikov and biweight kernels allot.

Triweight. Data close to the current point receive higher weights than any other kernel allows. Extreme cases get very little weight.

Gaussian. Weights follow a normal distribution, resulting in higher weighting of extreme cases than the Epanechnikov, biweight, tricube, and triweight kernels.

Cauchy. Extreme values receive more weight -- more than the other kernels, with the exception of the uniform, allow.

Smoothing Methods

The smoothing method determines how the kernel-weighted points are combined into an estimate. Select one of the following methods:

Mean. The arithmetic average of the weighted points.

Trim. The mean of the weighted points after discarding the most extreme 50% in the current region.

Median. The median of the weighted points.

Polynomial. The polynomial regression estimate at the current point. Select either a linear, quadratic, or cubic polynomial by specifying a degree of 1, 2, or 3.

Robust. A polynomial regression estimate resistant to outliers. Select either a linear, quadratic, or cubic polynomial by specifying a degree of 1, 2, or 3.

Using Commands

After selecting a data file with USE, continue with:

```
SMOOTH
   USE DATAFILE
   MODEL dep = pred1 pred2
   SAVE outfile
   ESTIMATE / WINDOW = FIXED
                       KNN,
              KERNEL = UNIFORM
                       EPANECHNIKOV
                       BIWEIGHT
                       TRIWEIGHT
                       TRICUBE
                       GAUSSIAN
                        CAUCHY
              SMOOTHER = MEAN
                       TRIM
                       MEDIAN
                       POLY
                        ROBUST
              BANDWIDTH = b
              NEIGHBORS = n
              PROPORTION = p
              GRID = n
              CANONICAL
              LOESS
              SAMPLE = BOOT(m, n)
                     = SIMPLE(m, n)
                     = JACK
```

Usage Considerations

Types of data. Smoothing requires a rectangular data file.

Print options. The output is standard for all PLENGTH options.

Quick Graphs. If a grid has been defined, the plot displays the data with the smoothing line for two-dimensional smoothing and creates a contour plot for three-dimensional smoothing. If no grid points are defined, the plot contains the residuals versus the smoothed values.

Saving files. SMOOTH saves predicted values and residuals if the number of grid points equals 0. Otherwise, the saved file contains grid points and smooth estimates.

BY groups. SMOOTH performs separate analyses for each level of any BY variables.

Case frequencies. You can use a FREQUENCY variable to duplicate cases.

Case weights. Smoothing uses a WEIGHT variable, if present, to weight cases.

Examples

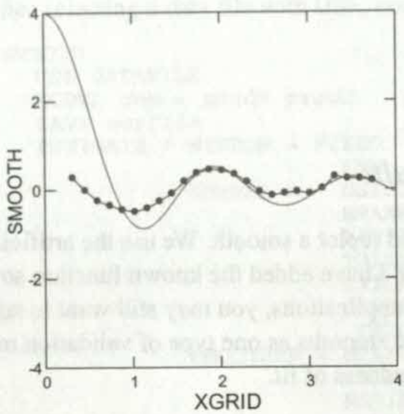
Example 1
Smoothing: Saving and Plotting Results

This example shows how to save and replot a smooth. We use the artificial data set from the introduction to this chapter. I have added the known function so that direct comparison is possible. In real data applications, you may still want to superimpose parametric models on nonparametric smooths as one type of validation method to supplement more formal tests of goodness of fit.

The input is:

```
NEW
RSEED 1357
REPEAT 150
LET X=SQR(CASE/10)
LET Y=SIN(4*X)/X+ZRN()/2+1.5*ZRN()*(URN()>.8)
ESAVE DAMP1
USE DAMP1
SMOOTH
MODEL Y=X
SAVE TEMP
ESTIMATE / LOESS,DEGREE=2
USE TEMP
BEGIN
PLOT SMOOTH*XGRID / LINE,XMIN=0,XMAX=4,
                    YMIN=-4,YMAX=4,FILL
FPLOT Y=SIN(4*X)/X ; XMIN=0,XMAX=4,YMIN=-4,YMAX=4,
                     AXES=0,SCALE=0,COLOR=RED
END
```

The output is:



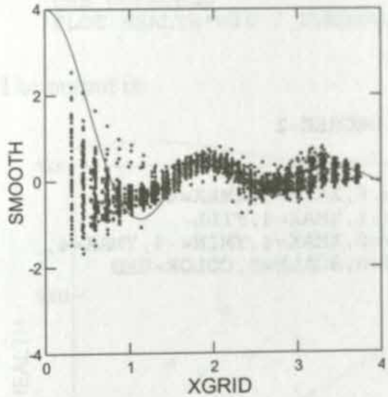
Example 2
Confidence Intervals for Smoothers

Confidence intervals on nonparametric smoothers can often be difficult to construct. One way around this problem is to bootstrap a smoother and examine the estimated values to get an idea of the shape of the envelope for a particular number of samples. The following example computes 100 bootstrap replications of the fit in the "Smoothing: Saving and Plotting Results" example. We have superimposed the theoretical curve on the sample results.

The input is:

```
SMOOTH
  GRAPH NONE
  USE DAMP1
  MODEL Y=X
  SAVE TEMP
  RSEED 1357
  ESTIMATE / LOESS,GRID=25,DEGREE=2,SAMPLE=BOOT(100)
  USE TEMP
  BEGIN
  PLOT SMOOTH*XGRID / SIZE=.5,XMIN=0,XMAX=4,
                      YMIN=-4,YMAX=4,FILL
  FPLOT Y=SIN(4*X)/X ; XMIN=0,XMAX=4,YMIN=-4,YMAX=4,
                       AXES=0,SCALE=0,COLOR=RED
  END
  GRAPH
```

The output is:



Notice that the estimates fail to envelop the leftmost segment of the curve. There are not enough points in the original sample to capture the variation in this area. This is, in part, a consequence of the original sampling scheme we chose, which produced fewer points at small *x* values than at larger.

How good are these bootstrap intervals? One way to find out in this case is to take 100 samples from the population and fit a LOESS to each sample. The following program generates 100 DAMP data sets. All 100 data sets are grouped together in a single file of 15000 cases, with an additional grouping variable G to denote the sample number. Because we reset the random number seed to the value used in the "Smoothing: Saving and Plotting Results" example, the first sample is identical to the data used in that example.

The input is:

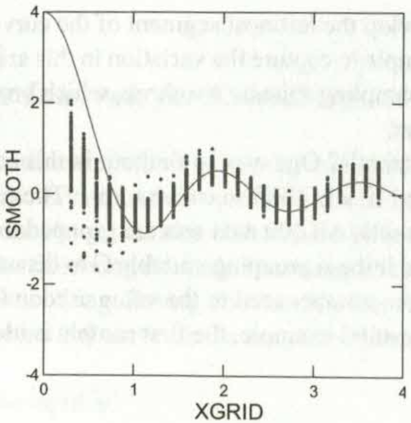
```
NEW
RSEED 1357
REPEAT 15000
LET X=SQR((1+MOD((CASE-1),150))/10)
LET Y=SIN(4*X)/X+ZRN()/2+1.5*ZRN()*(URN()>.8)
LET G=1+INT((CASE-1)/150)
ESAVE DAMP2
```

The following program runs 100 different LOESS smoothings and saves the results into a single file. Then we plot the results as before.

The input is:

```
GRAPH NONE
USE DAMP2
BY G
SMOOTH
MODEL Y=X
SAVE TEMP
ESTIMATE / LOESS,GRID=25,DEGREE=2
USE TEMP
BEGIN
PLOT SMOOTH*XGRID / SIZE=.5,XMIN=0,XMAX=4,
                    YMIN=-4,YMAX=4,FILL
FPLOT Y=SIN(4*X)/X ; XMIN=0,XMAX=4,YMIN=-4,YMAX=4,
                     AXES=0,SCALE=0,COLOR=RED
END
GRAPH
```

The output is:



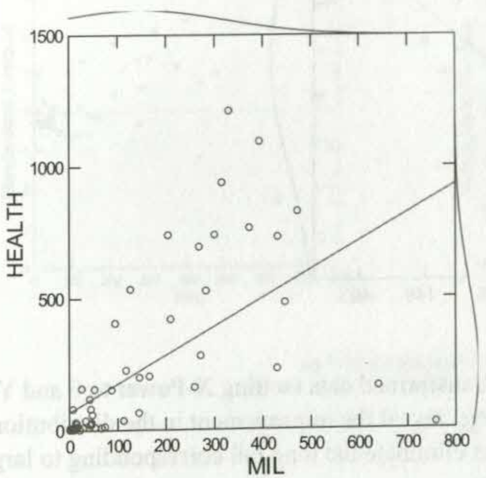
Example 3
Polynomial Regression and Smoothing

In this example, we illustrate the correspondence between polynomial smoothing in scatterplots and in the SMOOTH procedure. We also compare nonparametric smooths with their parametric counterparts. Throughout this example, we focus on the relationship between military and health spending.

The input is:

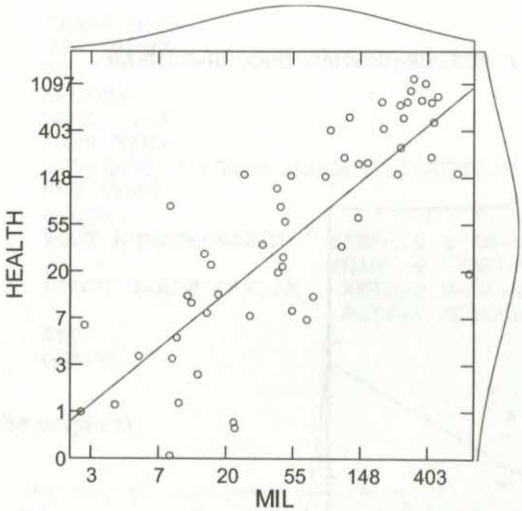
```
USE OURWORLD
PLOT HEALTH*MIL / BORDER=NORMAL SMOOTH=LINEAR
```

The output is:



Notice the large number of observations in the lower left corner of the plot. The border plots reveal a heavy concentration of cases at the low end of each axis and a few cases at the high end. Data of this type are not well suited for linear regression. However, transforming the data may improve the situation. Use the Graph Properties Dialog to investigate the effects of transforming both variables.

The plot resulting from applying the log transformation to both axes follows:



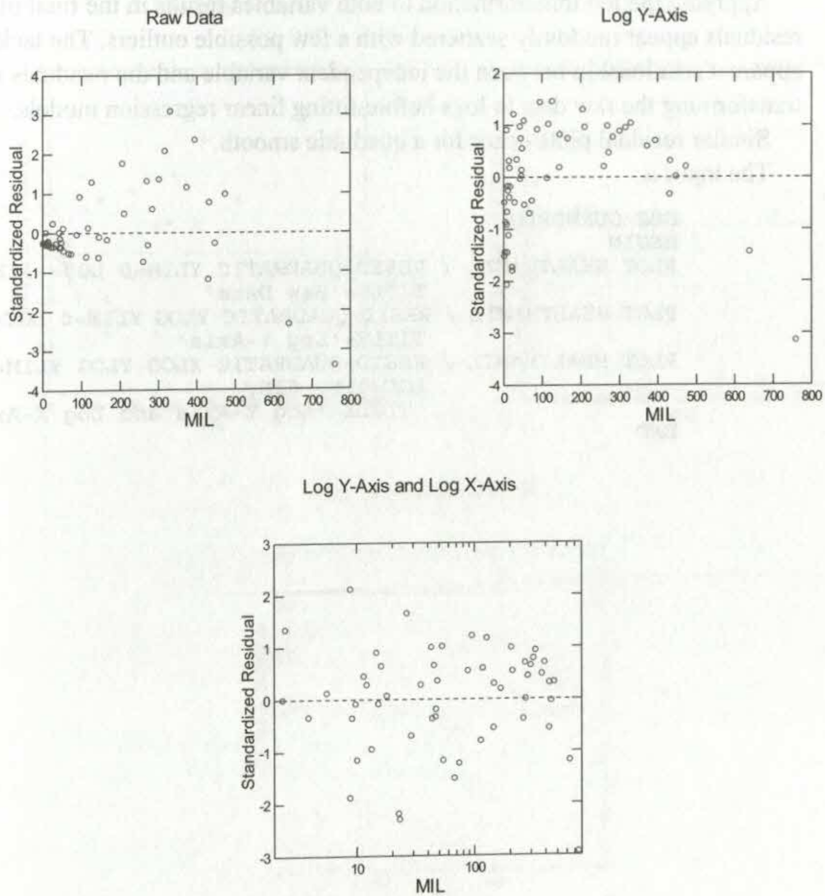
The border plots for the log transformed data (setting X-Power to 0 and Y-Power to 0 in the Graph Properties Dialog) reveal the improvement in the distributions for both variables; the transformations eliminate the long tail corresponding to large values of spending for each distribution.

Let's look at the residuals before and after log transforming the data.

The input is:

```
USE OURWORLD
BEGIN
PLOT HEALTH*MIL / RESID=LINEAR YLIM=0 LOC=-3IN,0IN,
                  TITLE='Raw Data'
PLOT HEALTH*MIL / RESID=LINEAR YLOG YLIM=0 LOC=3IN,0IN,
                  TITLE='Log Y-Axis'
PLOT HEALTH*MIL / RESID=LINEAR XLOG YLOG YLIM=0,
                  LOC=0IN,-6IN,
                   TITLE='Log Y-Axis and Log X-Axis'
END
```

The output is:



The residual plot for the raw data displays a pattern commonly referred to as a *right-opening megaphone*. As military spending increases, the residuals tend to get larger. This pattern often occurs when the dependent variable varies over a large range (from .385 to 1209.077, in this example). One common procedure used to alleviate the resulting nonconstant variance involves applying the log transformation to the dependent variable.

In the second plot, only the dependent variable (*HEALTH*) is log transformed. The plot displays a distinct nonlinear relationship; the extreme *MIL* values correspond to

negative residuals and values between 100 and 400 yield positive residuals. To control for this pattern, transform the independent variable or use a nonlinear model.

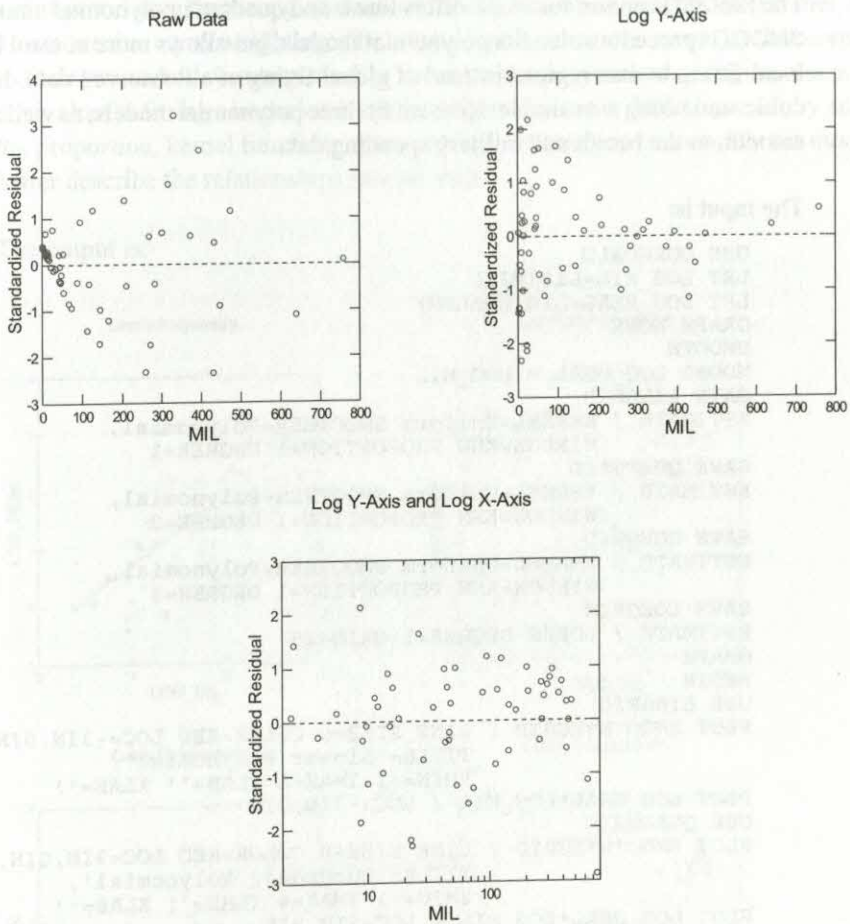
Applying the log transformation to both variables results in the final plot. The residuals appear randomly scattered with a few possible outliers. The lack of an apparent relationship between the independent variable and the residuals suggests transforming the raw data to logs before fitting linear regression models.

Similar residual plots occur for a quadratic smooth.

The input is:

```
USE OURWORLD
BEGIN
PLOT HEALTH*MIL / RESID=QUADRATIC YLIM=0 LOC=-3IN,0IN,
                  TITLE='Raw Data'
PLOT HEALTH*MIL / RESID=QUADRATIC YLOG YLIM=0 LOC=3IN,0IN,
                  TITLE='Log Y-Axis'
PLOT HEALTH*MIL / RESID=QUADRATIC XLOG YLOG YLIM=0,
                  LOC=0IN,-6IN,
                   TITLE='Log Y-Axis and Log X-Axis'
END
```

The output is:



The first plot displays a right-opening megaphone. In contrast to the linear smooth, however, transforming the dependent variable yields decreasing residual variance as *MIL* increases (a left-opening megaphone). Transforming both variables results in a random scatter of residuals.

Polynomial Smoothing

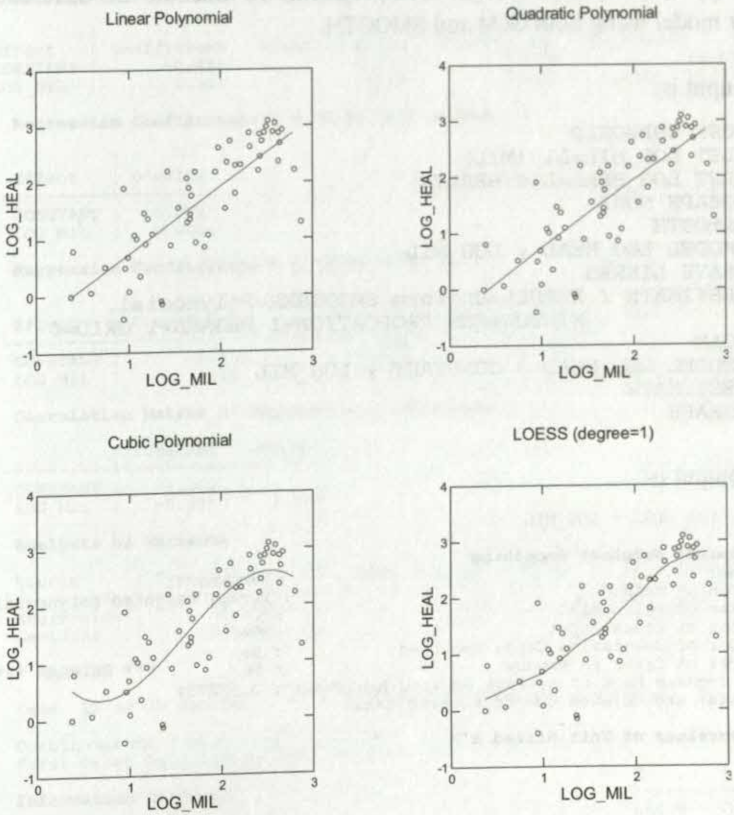
The SMOOTH option for PLOT offers linear and quadratic polynomial smoothing. The SMOOTH procedure also fits polynomial models, but allows more control by allowing local fitting to data regions instead of global fitting of all observed data. In addition, cubic smoothing is available. Here we fit three polynomial models, as well as a LOESS smooth, to the health and military spending data.

The input is:

```
USE OURWORLD
LET LOG_MIL=L10(MIL)
LET LOG_HEAL=L10(HEALTH)
GRAPH NONE
SMOOTH
MODEL LOG_HEAL = LOG_MIL
SAVE LINGRID
ESTIMATE / KERNEL=Uniform SMOOTHER=Polynomial,
           WINDOW=KNN PROPORTION=1 DEGREE=1
SAVE QUADGRID
ESTIMATE / KERNEL=Uniform SMOOTHER=Polynomial,
           WINDOW=KNN PROPORTION=1 DEGREE=2
SAVE CUBGRID
ESTIMATE / KERNEL=Uniform SMOOTHER=Polynomial,
           WINDOW=KNN PROPORTION=1 DEGREE=3
SAVE LOEGRID
ESTIMATE / LOESS DEGREE=1 GRID=25
GRAPH
BEGIN
USE LINGRID
PLOT SMOOTH*XGRID / LINE SIZE=0 COLOR=RED LOC=-3IN,0IN,
                    TITLE='Linear Polynomial',
                    YMIN=-1 YMAX=4 YLAB='' XLAB=''
PLOT LOG_HEAL*LOG_MIL / LOC=-3IN,0IN
USE QUADGRID
PLOT SMOOTH*XGRID / LINE SIZE=0 COLOR=RED LOC=3IN,0IN,
                    TITLE='Quadratic Polynomial',
                    YMIN=-1 YMAX=4 YLAB='' XLAB=''
PLOT LOG_HEAL*LOG_MIL / LOC=3IN,0IN
USE CUBGRID
PLOT SMOOTH*XGRID / LINE SIZE=0 COLOR=RED LOC=-3IN,-6IN,
                    TITLE='Cubic Polynomial',
                    YMIN=-1 YMAX=4 YLAB='' XLAB=''
PLOT LOG_HEAL*LOG_MIL / LOC=-3IN,-6IN
USE LOEGRID
PLOT SMOOTH*XGRID / LINE SIZE=0 COLOR=RED LOC=3IN,-6IN,
                    TITLE='LOESS (degree=1)',
                    YMIN=-1 YMAX=4 YLAB='' XLAB=''
PLOT LOG_HEAL*LOG_MIL / LOC=3IN,-6IN
END
```

By default, SMOOTH generates estimates at 25 grid points. For the linear, quadratic, and cubic models, we set the proportion of neighbors included when generating each estimate to 1 resulting in a global smooth of all data. By using a uniform kernel, all data points receive equal weighting. These specifications result in smooths that correspond to their parametric counterparts. For example, the polynomial of degree one is equivalent to fitting a linear least-squares regression to the data. However, by adjusting the proportion, kernel function, or smoothing method, you can modify the model to better describe the relationships in your data.

The output is:



The linear and quadratic smooths are practically indistinguishable. The cubic smooth begins turning up at the left end and turning down at the right end. Cubic models often

respond dramatically to data at the extremes of the observed range. In contrast, the LOESS smooth merely levels off slightly at the upper end in response to the two values at the right extreme. It appears that a first degree polynomial describes the relationship between the two transformed variables adequately; polynomials of higher degree overfit this data.

Mean Squared Error in Smoothing

When the number of grid points equals 0, SMOOTH generates a smooth estimate for each case and reports an MSE value. However, this mean squared error differs from the MSE typically reported in regression programs. To illustrate this difference, we fit a linear model using both GLM and SMOOTH.

The input is:

```
USE OURWORLD
LET LOG_MIL=L10(MIL)
LET LOG_HEAL=L10(HEALTH)
GRAPH NONE
SMOOTH
MODEL LOG_HEAL = LOG_MIL
SAVE LINRES
ESTIMATE / KERNEL=Uniform SMOOTHER=Polynomial,
           WINDOW=KNN PROPORTION=1 DEGREE=1 GRID=0
GLM
MODEL LOG_HEAL = CONSTANT + LOG_MIL
ESTIMATE
GRAPH
```

The output is:

```
Model: LOG_HEAL = LOG_MIL

  K-nearest Neighbor Smoothing
  Kernel                                            : Uniform
  Smoothing Method                                  : Kernel Weighted Polynomial Regression
  Degree of Polynomial                              : 1
  Number of Cases Input                             : 57
  Number of Nonmissing Cases Smoothed               : 56
  Number of Cases in Window                         : 56
  MSE (Square Root of Average Squared Residual)     : 0.572734
1 case(s) are deleted due to missing data.

  Eigenvalues of Unit Scaled X'X

     1      2
 --------------
  1.937   0.063
```

```
Condition Indices

  1        2
--------------
1.000    5.564

Variance Proportions

          :    1        2
----------+----------------
CONSTANT  : 0.031    0.969
LOG_MIL   : 0.031    0.969

Dependent Variable              : LOG_HEAL
N                               : 56
Multiple R                      : 0.808
Squared Multiple R              : 0.653
Adjusted Squared Multiple R     : 0.646
Standard Error of Estimate      : 0.583
```

Regression Coefficients $B = (X'X)^{-1}X'Y$

```
          :                                       Std.
Effect    : Coefficient   Standard Error   Coefficient   Tolerance      t
CONSTANT  :    -0.451          0.224          0.000          .       -2.016
LOG_MIL   :     1.191          0.118          0.808       1.000      10.073
```

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

```
          :
Effect    : p-value
----------+---------
CONSTANT  :   0.049
LOG_MIL   :   0.000
```

Regression Coefficients $B = (X'X)^{-1}X'Y$

```
          :                   95.0% Confidence Interval
Effect    : Coefficient                                          VIF
----------+------------------------------------------------------------
CONSTANT  :   -0.451         -0.900             -0.002
LOG_MIL   :    1.191          0.954              1.428       1.000

Correlation Matrix of Regression Coefficients

          : CONSTANT   LOG_MIL
----------+--------------------
CONSTANT  :   1.000
LOG_MIL   :  -0.937     1.000

Analysis of Variance

Source      : Type III SS   df   Mean Squares   F-ratio   p-value
------------+-------------------------------------------------------
Regression  :    34.516      1       34.516     101.467     0.000
Residual    :    18.369     54        0.340

*** WARNING *** :

Case  22 is an Outlier            (Studentized Residual : -3.281 )

Durbin-Watson D Statistic         : 1.472
First Order Autocorrelation       : 0.248

Information Criteria

AIC                : 102.500
AIC (Corrected)    : 102.961
Schwarz's BIC      : 108.576
```

The mean-square error (residual) from the ANOVA table for the linear model is 0.340. But the smoothing output reports a MSE value of 0.573 for this model. How is this possible?

To calculate the MSE for smoothing, SYSTAT squares the residuals, determines the average of these squared terms, and reports the square root of this average. The following commands compute the squared MSE value for the linear model.

The input is:

```
USE LINRES
LET SQRES=RESIDUAL*RESIDUAL
CSTATISTICS SQRES / N SUM MEAN
```

The output is:

```
                  ¦  SQRES
------------------+-------
N of Cases        ¦ 56.000
Sum               ¦ 18.369
Arithmetic Mean   ¦  0.328
```

The reported mean equals 18.369 / 56. Taking the square root of the mean:

```
CALC SQR(.328)
```

results in the reported MSE value of .573.

Notice that the sum of the squared residuals from SMOOTH equals the sum-of-squares (SS) for Residual in GLM. In GLM, the MS for error equals the SS for error divided by the degrees of freedom for error. In this case, 18.369 / 54 = 0.340. Thus, the two procedures differ in divisors for the sum of squared residuals (56 vs. 54). In the regression approach, we divide by (the number of cases - the number of estimated parameters). In the nonparametric smoothing approach, no parameters are being estimated, so we simply divide by the number of cases. Furthermore, the MSE in GLM is a sum of squares divided by a divisor. In SMOOTH, the MSE equals the *square root* of a sum of squares divided by a divisor.

Example 4
Smoothing Binary Data in Three Dimensions

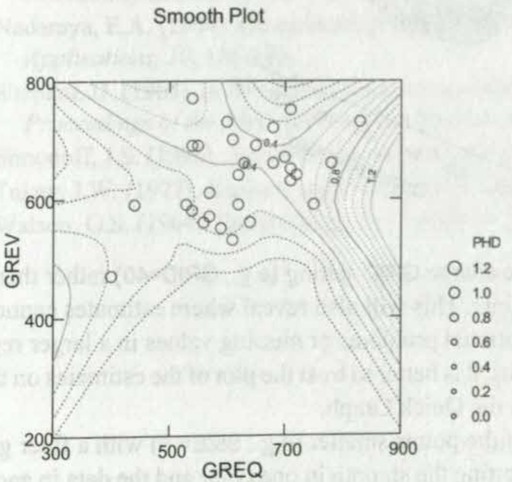
This example shows a smooth of binary data using two predictors: Graduate Record Examination Verbal and Quantitative scores. The dependent variable is a binary

indicator of whether or not a student was awarded a Ph.D. in a graduate psychology department. We employ the LOESS smoother.

The input is:

```
SMOOTH
   USE ADMIT
   MODEL PHD=GREV GREQ
   SAVE TEMP
   ESTIMATE / LOESS
```

The output is:



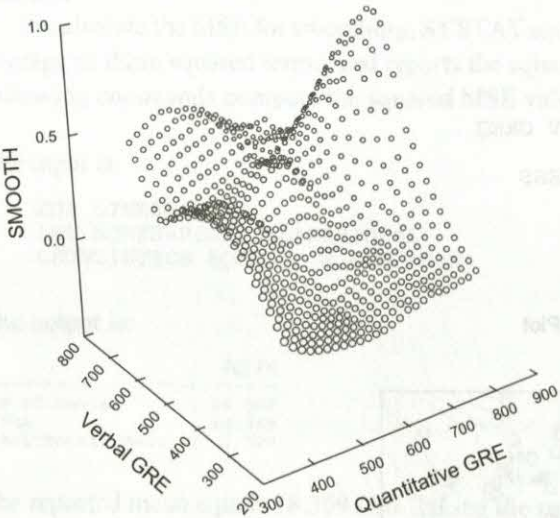
The Quick Graph uses a contour plot to represent the smoothed surface when there are two predictors. The size of the data points is proportional to the values on the dependent variable. The fitted surface is highest for Verbal GRE scores around 600 and Quantitative GRE scores around 700-800.

3D Scatter Plots

We plot the saved values as follows:

```
USE TEMP
PLOT SMOOTH*YGRID*XGRID / XLAB='Quantitative GRE',
                          YLAB='Verbal GRE'
```

The output is:



It is often preferable to use a large GRID setting (e.g., GRID=40) rather than to attempt to fit a surface to these points. This will also reveal where estimates cannot be made because there are computational problems or missing values in a larger region. Especially with 3D smooths, it is better to trust the plot of the estimates on the grid than the smoothed estimates in the Quick Graph.

You can set the SIZE of the points smaller (e.g., SIZE=.5) with a finer grid to make the surface less coarse. Plotting the smooth in one color and the data in another allows us to see both objects without either hiding the other.

References

Cleveland, W.S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cleveland, W.S. and Devlin, S. (1988). Locally weighted regression analysis by local fitting. *Journal of the American Statistical Association, 83*, 596-640.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, 2nd ed. London: Chapman & Hall.

Gonzalez, R.C. and Wintz, P. (1987). *Digital image processing*. Reading, MA: Addison-Wesley.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. London: Chapman and Hall.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge, UK: Cambridge University Press.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. London: Chapman and Hall.

McLain, D.H. (1974). Drawing contours from arbitrary data points. *The Computer Journal, 17,* 318-324.

Makridakis, S. and Wheelwright, S.C. (1989). *Forecasting methods for management*, 5th ed. New York: John Wiley & Sons.

Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics & Probability Letters, 7*, 195-199.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications, 10*, 186-190.

Shepard, D. (1965). A two-dimensional interpolation function for irregularly spaced data. *Proceedings of the 23rd National Conference of the ACM,* 517.

Simonoff, J.S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Series A, 26*, 359-372.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman and Hall.
Härdle, W. (1990). *Applied nonparametric regression*. Cambridge, U.K.: Cambridge University Press.
Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. London: Chapman and Hall.
McLain, D.H. (1974). Drawing contours from arbitrary data points. *The Computer Journal* 17, 318-324.
Makridakis, S. and Wheelwright, S.C. (1989). *Forecasting methods for management*. 5th ed. New York: John Wiley & Sons.
Marron, J.S. and Nolan, D. (1989). Canonical kernels for density estimation. *Statistics & Probability Letters* 7, 195-199.
Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186-190.
Shepard, D. (1968). A two-dimensional interpolation function for irregularly spaced data. *Proceedings of the 23rd National Conference of the ACM*, 517-523.
Simonoff, J.S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.
Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Series A*, 26, 359-372.

Chapter

12

Spatial Statistics

Leland Wilkinson

Spatial statistics compute a variety of statistics on a 2-D or 3-D spatially oriented data set. Variograms assist in the identification of spatial models. Kriging offers 2-D or 3-D kriging methods for spatial prediction. Simulation realizes a spatial model using Monte Carlo methods. Finally, a variety of point-based statistics are produced, including areas (volumes) of Voronoi polygons, nearest-neighbor distances, counts of polygon facets, and quadrat counts. Graphs are automatically plotted and summary statistics are printed for many of these statistics.

The geostatistical routines in SYSTAT Spatial are based on GSLIB (Deutsch and Journel, 1997). Point statistics are computed from a Voronoi/Delaunay partition of 2-D or 3-D configurations.

Resampling procedures are available in this feature.

Statistical Background

Spatial statistics involve a variety of methods for analyzing spatially distributed data. SYSTAT Spatial covers two principal areas: fixed-point methods (kriging and Gaussian simulation) and random-point methods (nearest-neighbor distances, polygon area/volumes, quadrat procedures). All of these procedures can be defined through a basic spatial model.

The Basic Spatial Model

The basic spatial model can be defined as follows. Assume $\mathbf{s}_0 \in \Re^d$ is a "site" or point in d-dimensional Euclidean space. The random variable $Z(\mathbf{s}_0)$ represents a

possible observation of some quantity or quality at the site $\mathbf{s}_0$. Instead of fixing $\mathbf{s}_0$ at a single site, however, we can let $\mathbf{s}$ vary over the index set $D \subset \Re^d$, so as to make $Z(\mathbf{s})$ a *stochastic process* or a *random field*:

$$\{Z(\mathbf{s}) : \mathbf{s} \in D\}$$

There are two principal variants of this random process. If D is a *fixed* subset of $\Re^d$, then we have a *geostatistical* model in which points are at predetermined locations and $Z(\mathbf{s})$ is sampled at these points. Although points are taken at fixed locations, the geostatistical model assumes $\mathbf{s}$ can vary continuously over $\Re^d$. On the other hand, if D is a *random* subset, then we have a *point process* in which $Z(\mathbf{s})$, or $\mathbf{s}$ itself, is sampled at random locations.

Conventional statistical procedures are unsuited for either of these models for several reasons. The major reason involves independence of observations. As with time series (see the SERIES module in SYSTAT), spatial models normally involve dependence among observations. We cannot assume that errors for a conventional statistical model applied to spatial data will be independent.

In the geostatistical model, the value of $Z(\mathbf{s}_i)$ is usually correlated with the value of $Z(\mathbf{s}_j)$. For example, if we sample groundwater level at a site, the value we find there can be predicted, in part, by the value at a nearby site. Furthermore, the nature of this relationship may be more complex than the one-dimensional dependencies found in a time series. The dependency structure may vary with direction, distance, and time. We expect nearby sites to be related. We might even find that groundwater level is related (perhaps negatively) to the level at a distant site and this relationship might vary over time.

In the point process model, we face a similar dependency issue even though sites are randomly distributed. We also encounter another problem: the statistics we construct in order to examine patterns and test hypotheses are not usually normal. The distribution of distances between pairs of points, the counts of points in fixed areas, the areas of clear space around points, and other spatial statistics are not normally distributed. So even if our interest is focused only on the distribution of sites (that is, we have no $Z(\mathbf{s})$ or $Z(\mathbf{s}) = \mathbf{s}$), we cannot resort to conventional statistical procedures.

We will first introduce the approaches designed to handle these problems in geostatistics and then summarize the basic approaches in point processes. The examples in the following sections should further highlight these issues.

The Geostatistical Model

The classic geostatistical model involves the random variate $Z(\mathbf{s})$ over a fixed field for $\mathbf{s}$ defined on the real numbers. The set

$$\{Z(\mathbf{s}) : \mathbf{s} \in D\}$$

where D is the collection of sites being studied, is a *random function* of the sites. The cumulative distribution function of $Z(\mathbf{s})$ is

$$F(\mathbf{s};z) = \mathrm{Prob}\{Z(\mathbf{s}) \le z\}$$

Fitting models and making inferences about $Z(\mathbf{s})$ requires us to make a global assumption about its behavior over all members of the set D. That is, summarizing $Z(\mathbf{s})$ usually requires using information from neighboring sites, and how that information is used depends on our global assumptions. These assumptions usually involve some form of *stationarity*. In its strong form, stationarity requires that

$$F(\mathbf{s}_1;z_1, \mathbf{s}_2;z_2, \ldots, \mathbf{s}_n;z_n) = F(\mathbf{s}_1 + \mathbf{h};z_1, \mathbf{s}_2 + \mathbf{h};z_2, \ldots, \mathbf{s}_n + \mathbf{h};z_n)$$

for all $\mathbf{s}_i \in D$, $\mathbf{s}_i + \mathbf{h} \in D$, $\mathbf{h} \in \Re^d$, and any finite n.

Because $\mathbf{h}$ acts as a translation vector, this condition is also called *stationarity under translation*. In geostatistical modeling, we often use a weaker form of the stationarity assumption:

$$E(Z(\mathbf{s})) = \mu \text{ for all } \mathbf{s} \in D\text{, and}$$

$$cov(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = C(\mathbf{s}_1 - \mathbf{s}_2) \text{ for all } \mathbf{s}_1, \mathbf{s}_2 \in D$$

The parameter μ is called the stationary mean.The function $C(.)$ is called a *covariogram*. These two conditions define weak, or *second-order stationarity* for $Z(\mathbf{s})$. The first implies that the mean is invariant over sites. The second implies that the covariance of random functions $Z(.)$ between all pairs of sites is a function of the difference between sites. Furthermore, if $C(.)$ is independent of direction (that is, it depends only on the Euclidean distance between $\mathbf{s}_1$ and $\mathbf{s}_2$), then we call it *isotropic*.

Variogram

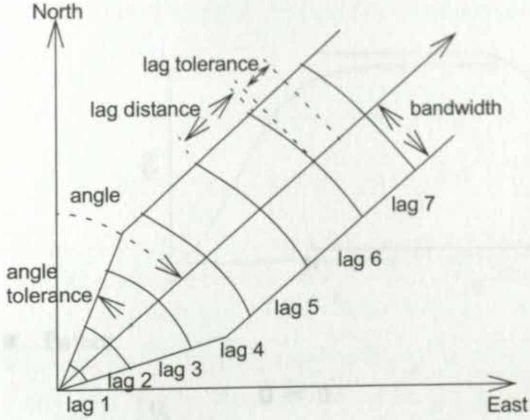
Instead of $C(.)$, geostatisticians usually work with a different but related function. This function is constructed from the variance of the first differences of the process:

$$2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = var(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))$$

The function $2\gamma(\mathbf{h})$, where $\mathbf{h}$ is the difference over all sites, is called the *variogram* and the function $\gamma(\mathbf{h})$ is called the *semi-variogram*. The classical estimator of the variogram function is:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N_{ij}} \sum_{i=1}^{n} \sum_{j<i} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2, \text{ where } \mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$$

When the data are irregularly spaced, the classical variogram estimator is computed with a "tolerance" region. The following figure shows the parameters for defining this region. The *angle* parameter determines the direction along which we want to compute the variogram. The *number of lags,* the *lag distance*, and the *lag tolerance* determine the maximum distance in this direction. The *bandwidth* determines the width of the band covering sites to be included in the calculations. And the *angle tolerance* determines the amount of tapering at the origin-end of the covering region. If this value is greater than 90 degrees, SYSTAT creates an omni-directional variogram, in which the full 360-degree sweep is used for computing lags. For three-dimensional spatial fields, these parameters are extended to the depth dimension from the usual horizontal (East) and vertical (North) dimensions.

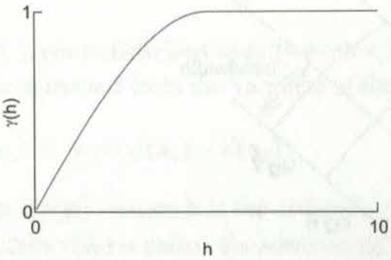


Variogram Models

We often want to construct a *variogram model* that fits our empirical variogram well. The smooth functions we use for variogram models not only help summarize the behavior of our process but they also give us a numerical method for fitting $Z(\mathbf{s})$ by least-squares. There are several popular functions for modeling the semi-variogram. The ones provided in SYSTAT (with scalar $h=\|\mathbf{h}\|$) are:

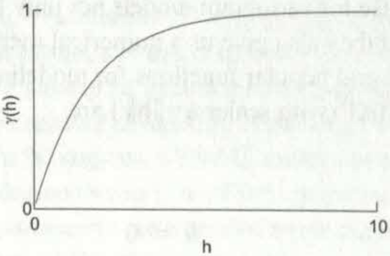
- **Spherical**

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + c\left[1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3\right], & 0 < h \le a \\ c_0 + c, & h > a \end{cases}$$



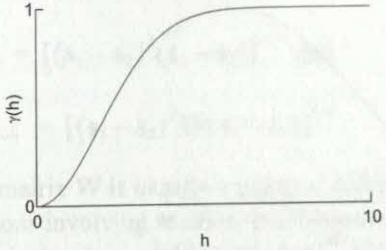
■ **Exponential**

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + c\left[1 - \exp\left(-\dfrac{3h}{a}\right)\right], & h > 0 \end{cases}$$



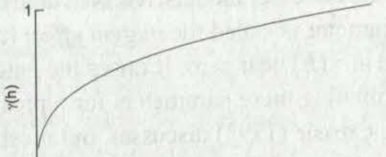
■ **Gaussian**

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + c\left[1 - \exp-\left(\dfrac{3h}{a}\right)^2\right], & h > 0 \end{cases}$$



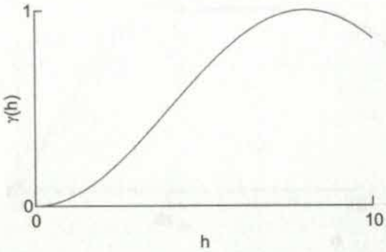
■ **Power**

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + ch^a, & h > 0,\ 0 < a < 2 \end{cases}$$



■ **Hole, or wave**

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + c\left[1 - \cos\left(\dfrac{\pi h}{a}\right)\right], & h > 0 \end{cases}$$



The hole model is sometimes parameterized equivalently with *sin* instead of *cos*. There is a variant of the hole effect model that includes a damping factor (usually exponential) for larger values of *h*. This model is not included in SYSTAT.

For most of these models, the parameter *c* is called the *sill*, and the parameter *a* is called the *range*. When appropriate, the sill is the maximum value of the function on the ordinate axis; in other cases, it is an asymptote. The range is the value of *h* for which $\gamma(h) = c_0 + c$. For other models, it is the value of *h* for which $\gamma(h)$ approaches $c_0 + c$. Another parameter is called the *nugget effect* (c_0). The nugget is an offset parameter measured at $\gamma(h)$ near zero. It raises the height of the entire curve (except for $\gamma(h) = 0$). Estimating these parameters for typical geophysical data presents some difficulties, as Cressie (1993) discusses, but most smoothing methods that use a variogram model are fairly robust against minor deviations in their values.

Variogram models can be combined in what Deutsch and Journel (1997) call the "nested model". This is a linear combination of submodels with separate parameter specifications for each. SYSTAT allows up to three submodels in a specification.

Pannatier (1996) offers an interactive program for variogram modeling (VARIOWIN) that is based on the same parameterizations as GSLIB and SYSTAT. Both VARIOWIN and SYSTAT offer variants of semi-variograms derived from GSLIB, such as the covariogram, correlogram, and semi-madogram. See Deutsch and Journel (1997) for details on these methods.

Anisotropy

If the variogram for a process is not identical for all directions, then it lacks isotropy. This *anisotropy* condition requires a more complex variogram model. The more basic type of anisotropy is *geometric*. This condition can be modeled by weighting distance according to direction, in a manner similar to the computation of Mahalanobis distance

in discriminant analysis (see the Discriminant procedure). That is, we compute $\gamma(h_w)$ instead of $\gamma(h)$, where

$$h = \|\boldsymbol{h}\| = [(\mathbf{s}_1 - \mathbf{s}_2)^T(\mathbf{s}_1 - \mathbf{s}_2)]^{1/2} \text{ and}$$

$$h_w = \|\mathbf{h}_w\| = [(\mathbf{s}_1 - \mathbf{s}_2)^T\mathbf{W}(\mathbf{s}_1 - \mathbf{s}_2)]^{1/2}$$

The weight matrix **W** is usually a positive definite composition of linear transformations involving rotation and dilatation. This turns the circular isometric locus for the isotropic model into an ellipse. SYSTAT specifies this matrix through several parameters. The first group specifies the angles for rotation: ANG1 is a deviation from north in a clockwise direction, ANG2 is a deviation from horizontal (for 3-D models), and ANG3 is a tilt angle. The second group specifies the shape of the ellipse that comprises a level curve for a given distance calculation: AHMAX is the maximum extent, AHMIN is the minimum extent, and AVERT is the 3-D (vertical) extent. An anisotropy index is calculated from these measures:

ANIS1=AHMIN/AHMAX and, for 3-D, ANIS2=AVERT/AHMAX

A second type of anisotropy is called *zonal*. This condition exists when different models apply to different directions. SYSTAT allows this type of modeling through nested variogram models. When the anisotropy parameter settings are different for each type of model in a nested structure, then we have a zonal isotropic model. See Journel and Huijbregts (1978) or Deutsch and Journel (1997) for further details.

Simple Kriging

The most popular geostatistical prediction method is called kriging, named after a South African mining engineer (Krige, 1962). Cressie (1990) provides a history of its origins in a variety of fields, including meteorology and physics. The simple kriging prediction model for $Z(\mathbf{s})$ is:

$$Z(\mathbf{s}) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) + \left(1 - \sum_{i=1}^{n} \lambda_i\right)\mu$$

where λ_i are weights and μ is the stationary mean. Kriging estimates weights λ_i that minimize the error variance over all estimated points $Z(\mathbf{s})$, not necessarily measured

at the given sites. The numerical estimation procedure requires a variogram model to be specified through the MODEL statement in SYSTAT.

Ordinary Kriging

By setting $\Sigma\lambda_i = 1$, we restrict the model and exclude the stationary mean. This constrained model is called *ordinary kriging*, the default method used in SYSTAT. The model is of the same form as that for simple kriging. Notice that, because the sum of the kriging weights is 1, the last summation term drops out of the model.

Universal Kriging

We may not be able to assume that $E(Z(\mathbf{s})) = \mu$ for all $\mathbf{s} \in D$, as we do in simple and ordinary kriging. Instead, we may want to assume that $E(Z(\mathbf{s}))$ is a linear combination of known functions $\{f_0(\mathbf{s}), ..., f_p(\mathbf{s})\}$. This allows us to model $Z(\mathbf{s})$ with trend components across the field. While these functions may be more general, it is customary to fit polynomial components to model this global trend in the following form:

$$Z(\mathbf{s}) = \sum_{j=0}^{p} \beta_j f_j(\mathbf{s}) + \delta(\mathbf{s})$$

The $\delta(\mathbf{s})$ term represents a stationary random process. SYSTAT offers linear and quadratic function terms for this type of modeling, including interactions. The terms are specified in the TREND command. Deutsch and Journel (1997) eschew the term "universal kriging" and instead call this method "kriging with a trend model." The SMOOTH=KRIG option of the PLOT command in SYSTAT is an independent implementation of universal kriging. This exploratory smoother does not offer the full modeling and output capabilities of the KRIG command in SPATIAL, however.

Simulation

Stochastic simulation offers the opportunity to create a realization of a spatial process to view the implications of a particular model or to estimate standard errors through Monte Carlo methods. Gaussian simulation generates the realization $\{z(\mathbf{s}):\mathbf{s} \in D\}$

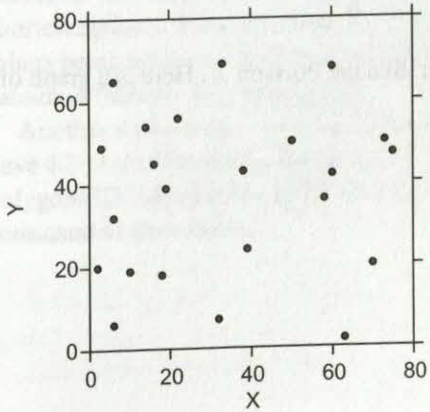
from the multivariate Gaussian $N(\mu, \Sigma)$, where the parameters for the centroid vector and covariance matrix are taken from the stationary means and covariances over the field.

SYSTAT implements the LUSIM algorithm from GSLIB (Deutsch and Journel, 1997). This algorithm requires the number of grid points and number of data points to be relatively small. It is designed to be most suited for a large number of realizations at a small number of nodes. SYSTAT executes one realization per use of the command, however, so the simulation is less useful for this purpose. To compensate, the memory requirements have been increased so that somewhat larger problems can be handled. See Deutsch and Journel (1997) and Haining (1990) for further details.

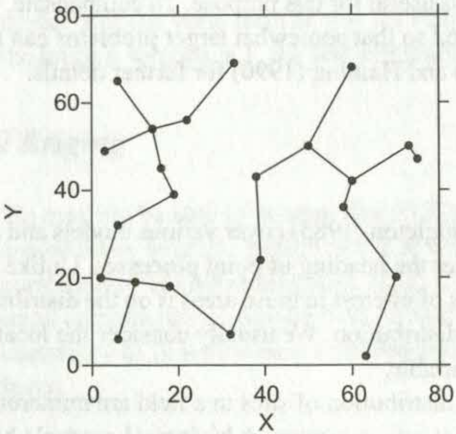
Point Processes

Cressie (1993) and Upton and Fingleton (1985) cover various models and applications that can be loosely grouped under the heading of point processes. Unlike traditional geostatistical methods, our focus of interest in these areas is on the distribution of sites themselves or functions of that distribution. We usually consider the location of sites in these cases to be a random variable.

The statistical indexes of the distribution of sites in a field are numerous. Most are based on some fundamental geometric measures. A biological example helps to illustrate these measures. The following plot shows the location of fiddler-crab holes in an 80 by 80 centimeter plot of the Pamet river marsh in Truro, Massachusetts (Wilkinson, 2005).



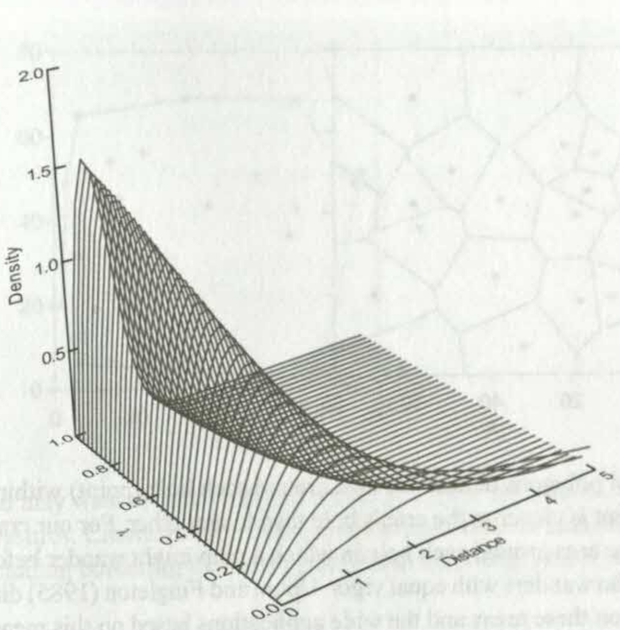
We might ask a number of questions about these data. First of all, are the holes randomly distributed in the plane? SYSTAT does not offer test statistics for answering this question directly but does provide the fundamental measures needed for a variety of such tests. The most widely used measure for spatial hypotheses of this sort is the nearest-neighbor distance, represented by the minimum spanning tree in the following figure (drawn with the SPAN option of the PLOT command in SYSTAT):



Upton and Fingleton (1985, Table 1.10) and Cressie (1993, Table 8.6) discuss a large number of statistical tests that are simple functions of these distances. The density of the nearest-neighbor distance under complete spatial randomness in two dimensions is:

$$g(d) = 2\pi\lambda d \exp(-\pi\lambda d^2), \quad d > 0$$

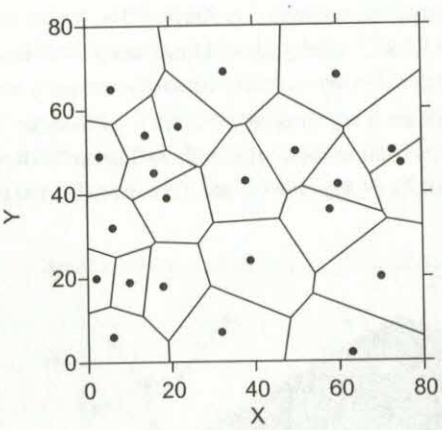
where λ is an intensity parameter, like the Poisson λ. Here is a graph of this density:



The density resembles the chi-square, with the shape evolving from normal to skewed as a function of the intensity parameter.

SYSTAT plots a histogram of these nearest-neighbor distances after the POINT command is issued. For our crabs, this distance has substantive meaning that can be useful for further modeling with other variables. The nearest-neighbor distance is the shortest distance from any crab hole to another in the sampled area. A crab is, all other things being equal, most likely to compete with the nearest-neighbor crab for local resources (absent remote foraging).

Another statistic used for tests of randomness is the Voronoi area, or volume if we have a 3-D configuration (flying crabs?). The following plot shows the Voronoi polygons (Dirichlet tiles) for the crab data. We used the VORONOI option of the PLOT command to draw these.

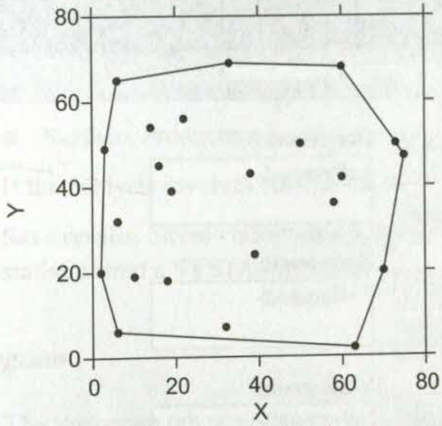


The Voronoi polygons delimit the area around each hole (point) within which every possible point is closer to the crab's hole than to any other. For our crabs, they might represent the area around each hole in which a crab might wander before hitting a neighbor who wanders with equal vigor. Upton and Fingleton (1985) discuss statistical tests based on these areas and the wide applications based on this measure. Okabe et al. (1992) cover Voronoi tessellations in depth.

A third statistic, also based on the Voronoi tessellation, is the count of the number of facets in each Voronoi polygon. For our crabs, this is a measure of the number of near neighbors each crab must contend with. It is positively correlated with the area measure, but is nevertheless distributed differently. Upton and Fingleton (1985) discuss applications.

A fourth measure of point intensity is the quadrat count. We simply count the number of points found in a set of rectangles defined by a grid (the SYSTAT GRID command). Upton and Fingleton (1985) discuss statistical tests based on this measure. Not surprisingly, several are chi-square based, following the rationale for using chi-square tests on binned one-dimensional variates.

Finally, Cressie (1993) and Upton and Fingleton (1985) discuss edge effects that can influence the distribution of many of these statistics. For example, the Voronoi areas (volumes) for points at the periphery of the configuration may be infinite or, because of the distribution of a few neighboring points, substantial outliers. These edge points also tend to have fewer neighbors as candidates for distance calculations. Consequently, it is often useful to be able to identify the points that lie on the convex hull in two or three dimensions. The following figure shows the hull for the 2-D crab data.



You may want to exclude points on the hull from analyses based on some of the above measures. Cressie (1993) discusses other methods for eliminating edge effects, including bordering the configuration and excluding points in the border.

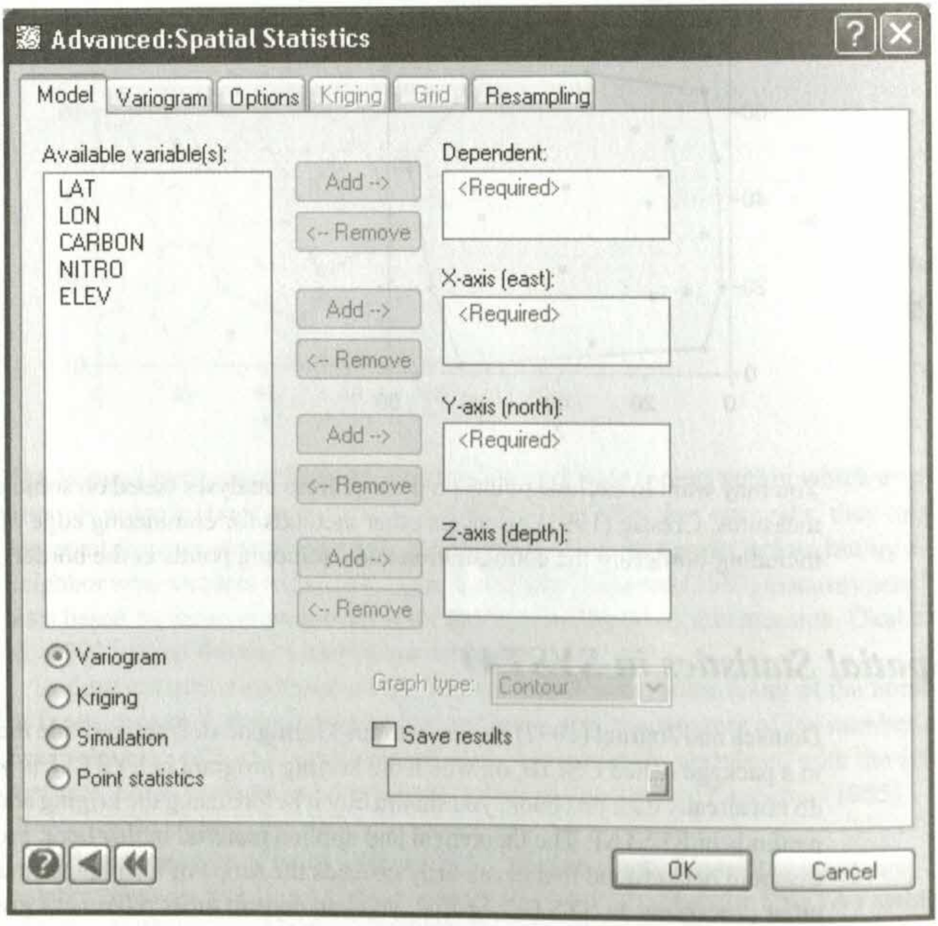
Spatial Statistics in SYSTAT

Deutsch and Journel (1997) discuss various kriging models and provide the algorithms in a package called GSLIB, on which the kriging program in SYSTAT is based. If you do not already own this book, you should buy it before using the kriging and simulation methods in SYSTAT. The theoretical and applied material in this book provides essential background that necessarily exceeds the scope of a computer manual. For other procedures in SYSTAT Spatial, you can consult other references given in this chapter.

Spatial Statistics Dialog Box

To open the Spatial Statistics dialog box, from the menus choose:

Advanced
  Spatial Statistics...



Specify the variables and select one of the following analyses:

- **Variogram.** Computes spatial dissimilarity measures over varying distances.
- **Kriging.** Generates predictions by minimizing the error variance over all estimated points.
- **Simulation.** Uses the multivariate normal distribution to generate a realization for a defined model. Simulation is often used to study a particular model or to estimate standard errors.
- **Point statistics.** Yields areas (volumes) of Voronoi polygons, nearest-neighbor distances, counts of polygon facets, and quadrat counts for sites by treating site locations as a random variable.

For Kriging and Simulation, select the graph type to display:

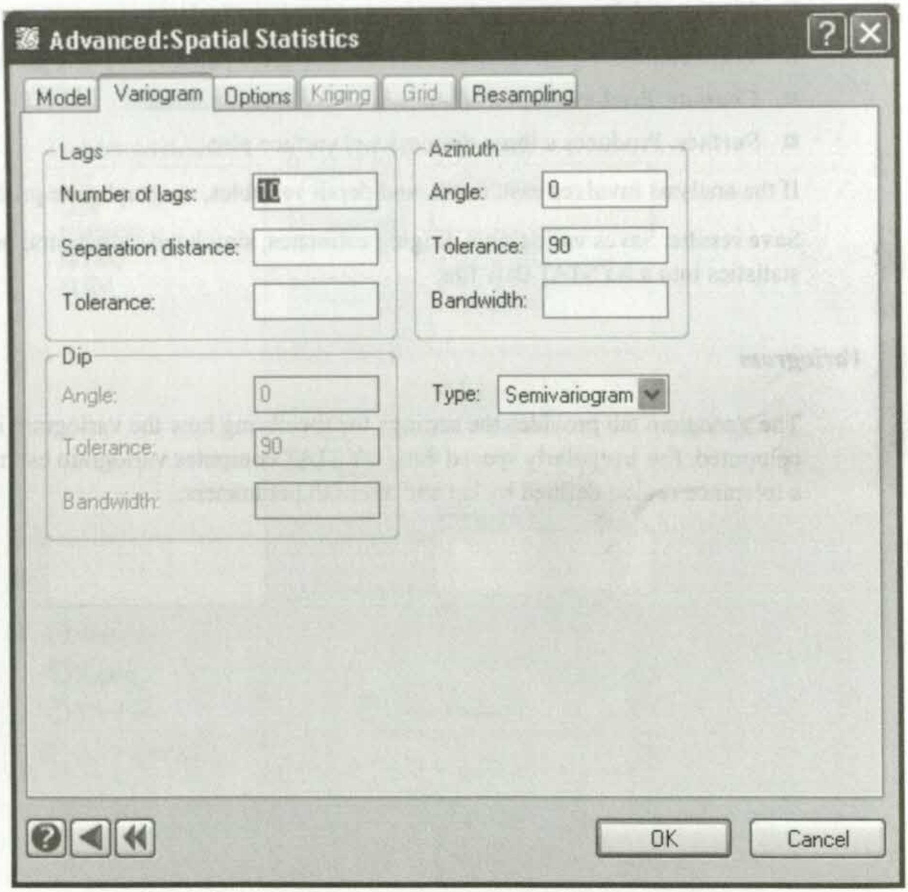
- **Tile.** Produces a plot contoured with shading fill patterns in color gradations.
- **Contour.** Produces a plot contoured with gradation lines.
- **Surface.** Produces a three-dimensional surface plot.

If the analysis involves east, north, and depth variables, no graphs are produced.

Save results. Saves variograms, kriging estimates, simulated values, and point statistics into a SYSTAT data file.

Variogram

The Variogram tab provides the settings for specifying how the variogram is to be computed. For irregularly spaced data, SYSTAT computes variogram estimates using a tolerance region defined by lag and azimuth parameters.



The lag parameters determine the maximum distance in the direction defined by the azimuth angle.

- **Number of lags.** Enter the number of lags used for calculating the spatial similarity measure.
- **Separation distance.** Specify the length of each lag.
- **Tolerance.** Specify a length to add to the separation distance to account for data on an irregular grid. This value is usually one-half of the separation distance (or smaller).

The azimuth parameters define the direction and width of the region used for the variogram.

- **Angle.** Defines the direction (in degrees clockwise from the North axis) along which the variogram is computed.
- **Tolerance.** Specify the amount of tapering (in degrees) near the origin for the covering region. For values exceeding 90 degrees, an omni-directional variogram results.
- **Bandwidth.** Specify the width of the band covering sites. Variogram calculations include points lying within the specified value in either direction from the vector defined by the azimuth angle.

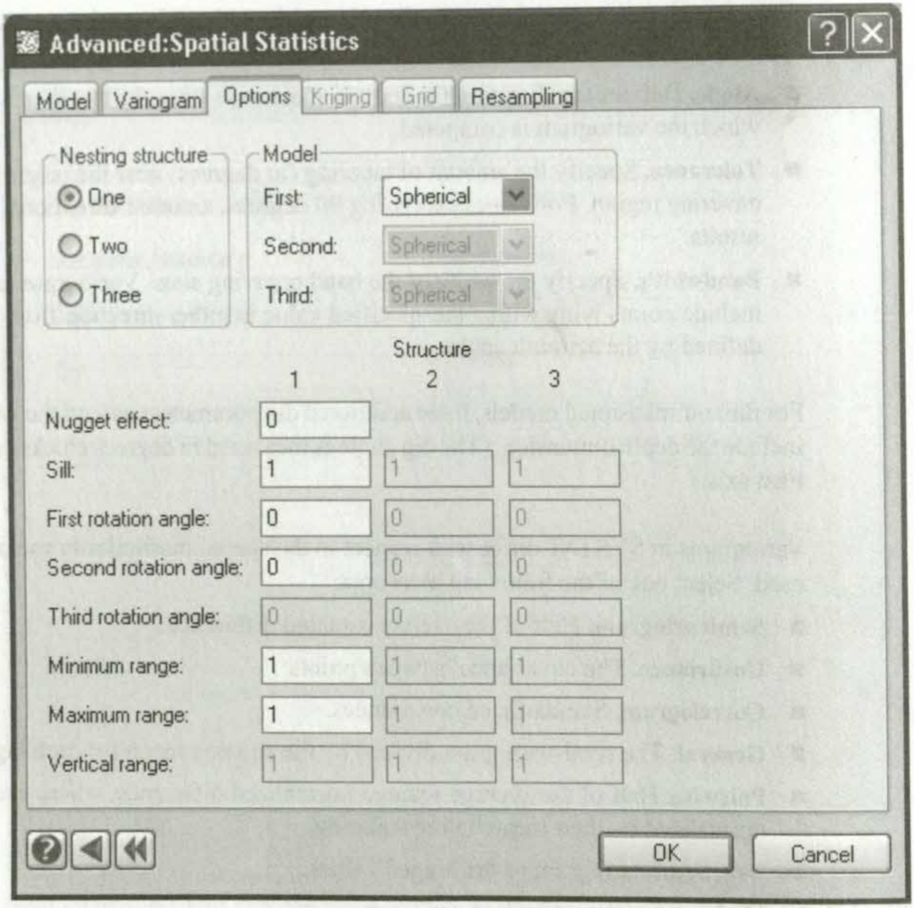
For three-dimensional models, three additional dip parameters extend the variogram to include the depth dimension. (The dip angle is measured in degrees clockwise from the East axis.)

Variograms in SYSTAT differ with respect to the spatial dissimilarity measure [γ(h)] used. Select one of the following measures:

- **Semivariogram.** Half of the average squared difference.
- **Covariance.** The covariance between points.
- **Correlogram.** Standardized covariances.
- **General.** The semi-variogram divided by the squared mean for each lag.
- **Pairwise.** Half of the average squared normalized difference, where each pair is normalized by their mean before squaring.
- **Log.** Semi-variogram of the logged values.
- **Madogram.** Mean absolute deviation.

Options

Define options specific to each analysis in the Options tab.The Model Options tab offers settings for defining the variogram model. Up to three nesting structures are allowed.



Nesting structure. Specify the number of nested structures.

For each structure, specify the form of the model. Alternatives include:

- **Spherical.** Near the origin, the spherical model is linear. The tangent to the curve at the origin reaches the sill at a distance of two-thirds of the distance at which the curve reaches the sill.
- **Exponential.** Near the origin, the exponential model is also linear. The tangent to the curve at the origin reaches the sill at a distance of one-third of the distance at which the value of the curve reaches 95% of the sill.
- **Gaussian.** Near the origin, this model is parabolic.

- **Power.** This model does not reach a sill. For exponents between 0 and 1, the model is concave; for values between 1 and 2, the model is convex. An exponent of 1 yields a linear variogram.
- **Hole.** The hole model oscillates around the sill.

In addition, specify the following:

Nugget effect. Enter the value at which the variogram intersects the vertical axis. Specifying a nugget raises the height of the variogram.

Sill. Enter the maximum value attained by the function. For some models, the sill is the asymptote of the function.

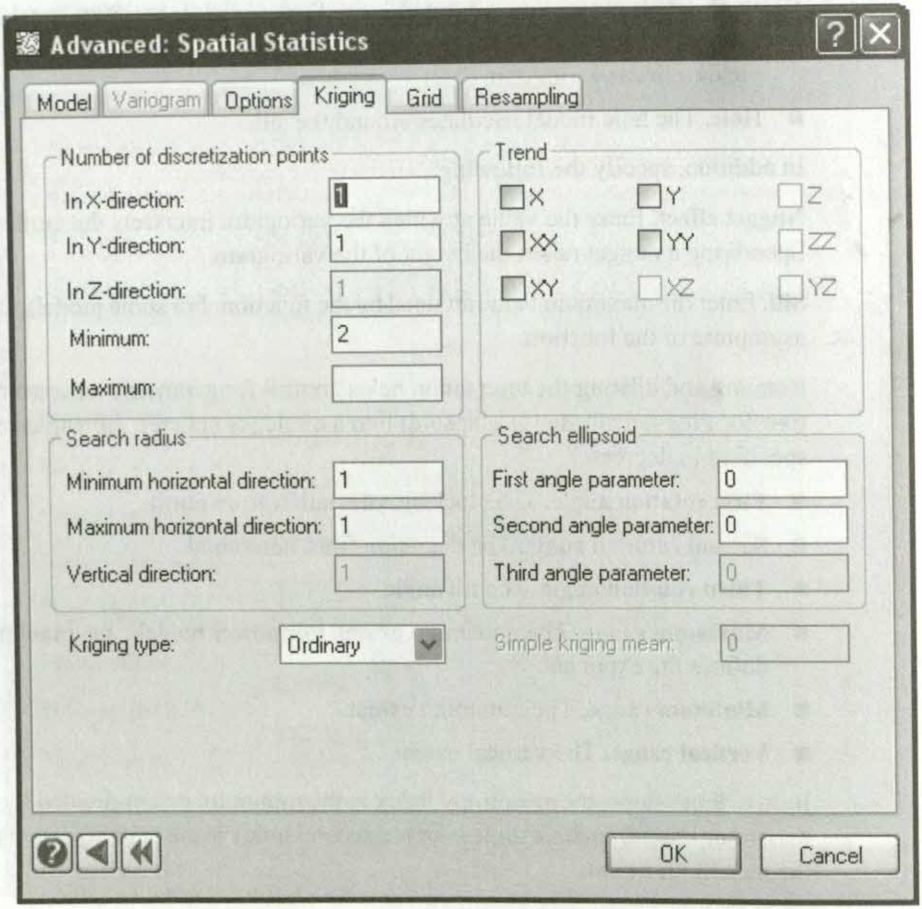
Rotating and dilating the orientation helps control for geometric anisotropy by transforming the ellipse (or ellipsoid) into a circle (or sphere). All angles must be specified in degrees.

- **First rotation angle.** The clockwise deviation from north.
- **Second rotation angle.** The deviation from horizontal.
- **Third rotation angle.** The tilt angle.
- **Maximum range.** The maximum extent. For power models, the maximum range defines the exponent.
- **Minimum range.** The minimum extent.
- **Vertical range.** The vertical extent.

In two dimensions, the anisotropy index is the minimum extent divided by the maximum extent. In three dimensions, a second index is the vertical extent divided by the maximum extent.

Kriging

Kriging yields estimates of the dependent variable based on nearby points, taking into account spatial relationships.



Number of discretization points. Specify the number of points for each block in the kriging analysis.

Trend. Defines polynomial trend components to add to the universal kriging analysis. X, Y, and Z correspond to the East, North, and Depth variables, respectively. For example, suppose the *x* axis (East) variable is *LONG*. Selecting XX adds the term *LONG*LONG* to the kriging model.

Search radius. Defines the size of the region used to compute the kriging estimates.

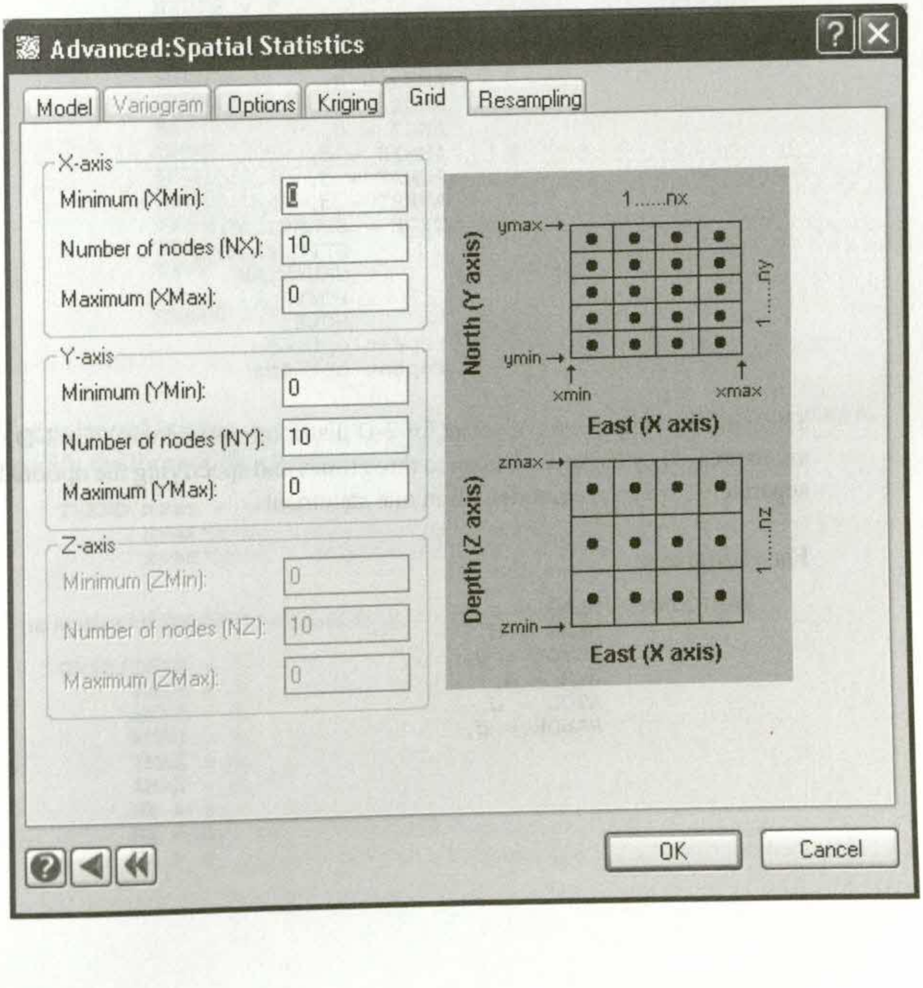
Search ellipsoid. Defines the orientation of the region used to compute the kriging estimates.

Three types of kriging are available:

- **Ordinary.** Constrains the sum of the kriging weights to be 1.
- **Simple.** Uses unconstrained weights. For simple kriging, specify the stationary mean.
- **Universal.** Kriging with polynomial trend components.

Grid

The Grid tab offers settings for determining the size and shape of the grid used for the kriging estimates and for quadrat counting in the point methods.



For each axis, specify:

- **Minimum.** The minimum value for the grid along the axis.
- **Number of nodes.** The number of points along the axis.
- **Maximum.** The maximum value along the axis.

SYSTAT uses equal spacing between consecutive nodes for each axis.

Using Commands

First, specify your data with USE FILENAME. Continue with:

```
SPATIAL
MODEL var = varlist / NUGGET = d,
                      SILL = d,
                      ANG1 = d,
                      ANG2 = d,
                      ANG3 = d,
                      AHMIN = d,
                      AHMAX = d,
                      AVERT = d,
                      TYPE = SPHERICAL
                             EXPONENTIAL
                             GAUSSIAN
                             POWER
                             HOLE,
                    / repeat options,
                    / repeat options
```

There are two arguments in varlist for 2-D distributions and three for 3-D. Submodels are expressed by using slashes up to three times and specifying the optional arguments separately for each submodel, all in one statement.

For variograms:

```
VARIOGRAM /NLAG = n,
           XLAG = d,
           XLTOL = d,
           AZM = d,
           ATOL = d,
           BANDH = d,
```

```
DIP = d,
DTOL = d,
BANDV = d,
TYPE = SEMI
       COVARIANCE
       CORRELOGRAM
       GENERAL
       PAIRWISE
       LOG
       MADOGRAM
```

For kriging:

```
KRIG / NXDIS = n,
       NYDIS n,
       NZDIS = n,
       NDMIN = d,
       NDMAX = d,
       RADMIN = d,
       RADMAX = d,
       RADVER = d,
       SANG1 = d,
       SANG2 = d,
       SANG3 = d,
       SKMEAN = d,
       TREND,
       TYPE = SIMPLE
              ORDINARY,
       GRAPH = CONTOUR
               TILE
               SURFACE
```

For universal kriging, use TYPE=ORDINARY and the TREND option. In addition, specify the form of the trend using the TREND command:

```
TREND xvar + yvar + zvar + ,
      xvar*xvar + yvar*yvar + zvar*zvar +
      xvar*yvar + xvar*zvar + yvar*zvar
```

The syntax of the GRID statement is:

```
GRID/XMIN = d,
     YMIN = d,
     ZMIN = d,
     XMAX = d,
     YMAX = d,
     ZMAX = d,
     NX = n,
     NY = n,
     NZ = n
```

The syntax of the SIMULATE and POINT statements follows:

```
SIMULATE/GRAPH = CONTOUR
                 TILE
                 SURFACE,
POINT varlist
```

Usage Considerations

Types of data. SPATIAL uses rectangular data only. The basis (spatial) variables are expected to be measures of latitude, longitude, depth, or other spatial dimensions. The dependent variable is expected to be symmetrically distributed or transformed to a symmetrical distribution.

Print options. There are no PLENGTH options. Output reports parameter settings. Graphs show the distributions and fitted models.

Quick Graphs. SPATIAL produces variograms, kriging surfaces, simulations, and nearest-neighbor histograms. You can choose the type of graph used to display the results of the KRIG and SIMULATE commands by using the GRAPH option.

Saving files. SPATIAL saves variograms, kriging estimates, simulated values, and point statistics.

BY groups. SPATIAL analyzes data BY groups. Your file need not be sorted on the BY variable(s).

Case frequencies. FREQ <variable> increases the number of cases by the FREQ variable.

Case weights. WEIGHT is not available in SPATIAL.

Examples

The examples begin with a kriging analysis of a spatial data set and proceed to simulation and point processes. Data in the "Point Statistics" example are used with the permission of Kooijman (1979).

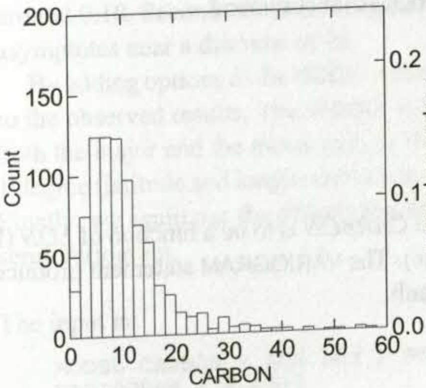
Example 1
Kriging (Ordinary)

The data in this example were taken from a compilation of worldwide carbon and nitrogen soil levels for more than 3500 scattered sites. These data were compiled by P. J. Zinke and A. G. Stangenberger of the Department of Forestry and Resource Management at the University of California, Berkeley. The full data set is available at the U.S. Carbon Dioxide Information Analysis Center (CDIAC) site on the World Wide Web. For our purposes, we have restricted the data to the continental U.S. and have averaged duplicate measurements at single sites by analyzing BY the *LAT* and *LON* variables using Basic Statistics and saving the averages.

The first step in the analysis is to examine the dependent variable (*CARBON*). The sample histogram for this variable is positively skewed.

```
USE SOIL
HIST CARBON
```

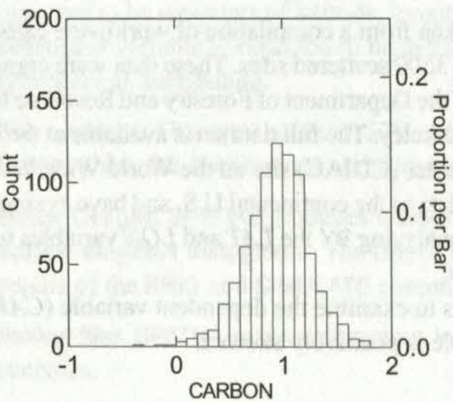
Here is the resulting histogram of the carbon levels:



We can use the the Axis tab of the Interactive Graphics Dialog box to transform the *CARBON* variable so that it looks more normally distributed. A value near 0 in the Power spin-box produces a normal-appearing histogram, suggesting that a log transformation can approximately symmetrize these data.

```
LET CARBON = L10(CARBON)
HIST CARBON
```

Here is the histogram for the log10 transformed data:



We will proceed using these log-transformed data.

The first step on the way to fitting a kriging surface is to identify a model through the variogram. We can get preliminary guidance for choosing a variogram model by using the default values of the VARIOGRAM command.

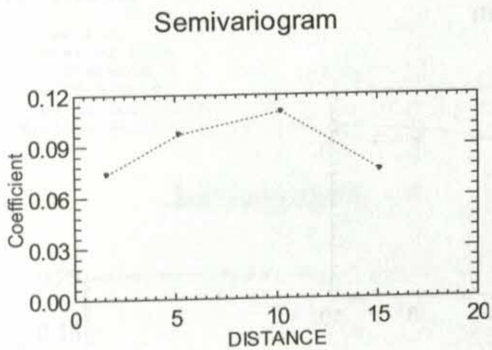
The input is:

```
SPATIAL
MODEL CARBON = LON LAT
VARIOGRAM
```

The MODEL statement specifies that *CARBON* is to be a function of *LON* (longitude of the sampling site) and *LAT* (latitude). The VARIOGRAM statement produces an omni-directional semi-variogram by default.

The output is:

```
Structural Model

  Nugget (c0)                                                            : 0.000
  First Rotation Angle (azimuth, or degrees clockwise from North)        : 0.000
  Second Rotation Angle (dip, or degrees down from azimuthal)            : 0.000
  First Anisotropy Index (anis1=ahmin/ahmax)                             : 1.000
  Sill (c)                                                               : 1.000
  Range (a)                                                              : 1.000

Semivariogram

  Direction                       :  0.000
  Number of Lags                  :     10
  Lag Distance                    :  5.630
  Lag Tolerance                   :  2.815
  Angular Tolerance               : 90.000
  Maximum Horizontal Bandwidth    :  5.630
```



The semi-variogram suggests several things. First, we need a nugget and sill value to offset our model semi-variogram high enough to reach a maximum value somewhere around 0.10. Second, we need to specify a range value so that the model variogram asymptotes near a distance of 10.

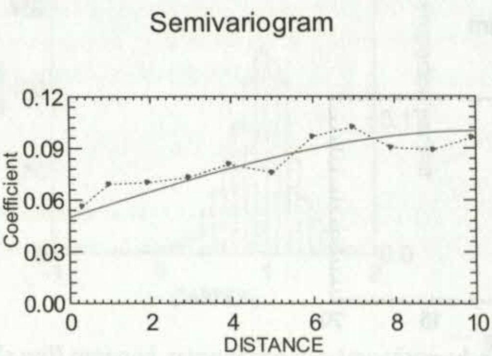
By adding options to the MODEL statement, we manage to fit a theoretical variogram to the observed results. The AHMAX and AHMIN parameters specify that the range for both the major and the minor axes is 10 degrees. We choose a lag distance (XLAG) of 1 degree (latitude and longitude) to base our variogram model on relatively local detail. Finally, we again use the default angular tolerance to produce an omni-directional semi-variogram.

The input is:

```
MODEL CARBON = LON LAT / SILL=.05,NUGGET=.05,AHMAX=10,AHMIN=10
VARIOGRAM / XLAG=1
```

The output is:

```
Structural Model
  Nugget (c0)                                                        : 0.050
  First Rotation Angle (azimuth, or degrees clockwise from North) : 0.000
  Second Rotation Angle (dip, or degrees down from azimuthal)     : 0.000
  First Anisotropy Index (anisl=ahmin/ahmax)                       : 1.000
  Sill (c)                                                           : 0.050
  Range (a)                                                          : 10.000

Semivariogram
  Direction                      :  0.000
  Number of Lags                 :    10
  Lag Distance                   :  1.000
  Lag Tolerance                  :  0.500
  Angular Tolerance              : 90.000
  Maximum Horizontal Bandwidth   :  1.000
```



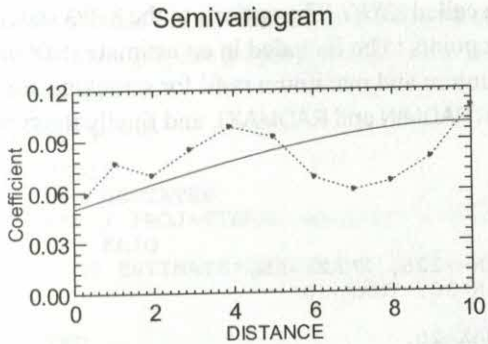
We can check whether we need to worry about anisotropy by checking semi-variograms for different angles. Here are two excursions, 90 degrees separated. By setting angular tolerance (ATOL) to 20 degrees, we keep the lagging window narrow at its origin-end.

The input is:

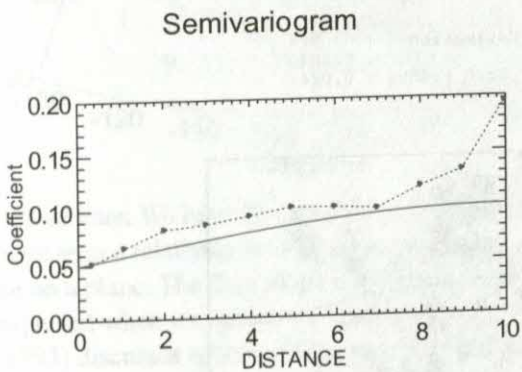
```
VARIOGRAM / XLAG=1,AZM=45,ATOL=20
VARIOGRAM / XLAG=1,AZM=135,ATOL=20
```

The output is:

```
Semivariogram

  Direction                      : 45.000
  Number of Lags                 :     10
  Lag Distance                   :  1.000
  Lag Tolerance                  :  0.500
  Angular Tolerance              : 20.000
  Maximum Horizontal Bandwidth   :  1.000
```




```
Semivariogram

  Direction                       : 135.000
  Number of Lags                  :      10
  Lag Distance                    :   1.000
  Lag Tolerance                   :   0.500
  Angular Tolerance               :  20.000
  Maximum Horizontal Bandwidth    :   1.000
```



There is enough difference to suggest the possibility of anisotropy, but not enough to radically alter the results. Some of this variety is likely due to the uneven scattering of the sites. This can induce negative or fluctuating spatial correlations, as evidenced by a wavy semi-variogram. If these are pronounced, one might consider a wave or hole-effect semi-variogram model.

We will keep our metric spherical nevertheless. This uneven scattering is not comforting, however; kriging normally should rest on fairly evenly distributed sampling sites or on a regular grid. Our data are received, however, so the sites are given.

Now we can use the spherical model to fit a surface to the carbon data. We add a GRID statement to specify the grid points where estimates are to be made. We also

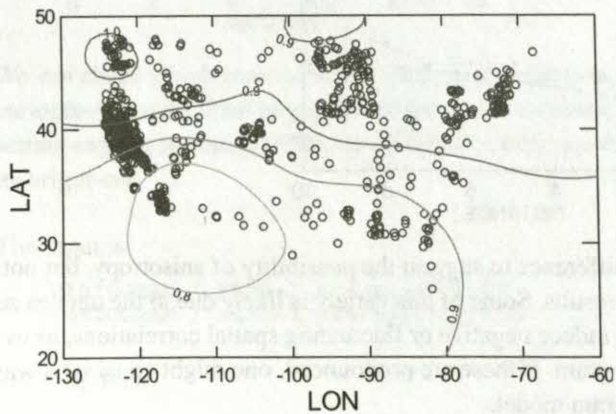
SAVE the estimates into a file called *KRIG*. The options to the KRIG statement specify the minimum number of data points to be included in an estimate (NDMIN), the maximum (NDMAX), the minimum and maximum radii for searching for neighboring sites to include in an estimate (RADMIN and RADMAX), and finally the type of graph (a CONTOUR plot).

The input is:

```
GRID /    NX=10, XMIN=-125, XMAX=-65,
          NY=10, YMIN=30, YMAX=50
SAVE KRIG
KRIG / NDMIN=2, NDMAX=20,
       RADMAX=5, RADMIN=5,
       GRAPH=CONTOUR
```

The output is:

```
Ordinary Kriging

  Search Radius1: 5.000
  Search Radius2: 5.000
  Search Angle 1: 0.000
  Search Angle 2: 0.000

  Number of Blocks used in Estimation      :     92
  Average Estimated Value                  : 0.882
  Number of Y (North-South) Grid Points    : 0.024
```

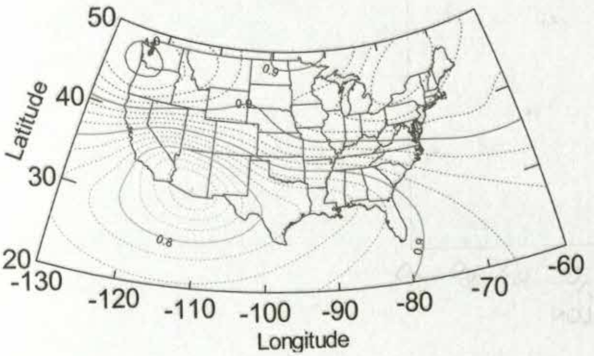


Our model shows the lowest soil carbon concentrations in the southwest and the highest in the north, particularly the northwest.

We can use our saved data to overlay these estimates on a map of the U.S. I have used a stereographic projection and set the axis limits to make the two graphs correspond.

The input is:

```
BEGIN
USE USSTATES
MAP / PROJ=STEREO,AX=4,SC=2,YMIN=20,YMAX=50
USE KRIG
PLOT ESTIMATE*GRID(2)*GRID(1)/PROJ=STEREO,CONTOUR,
                                YMIN=20,YMAX=50,SMOO=INVERSE,
                                AX=0,SC=0,SIZE=0,ZTICK=20

END
```



A final note: We have fit a surface to data distributed on the continental U.S. This represents a relatively small portion of the global sphere, so I have assumed the data to lie on a plane. The map projection makes clear that there is some distortion to be expected when we ignore the spherical nature of the coordinates, however. Cressie (1993) discusses spherical kriging methods, but they are not available in SYSTAT. Smaller areas, such as state, province, or county data, should be little cause for concern.

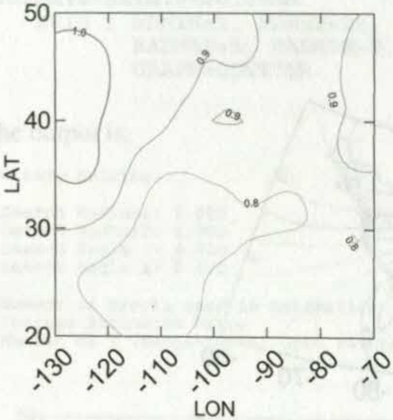
Example 2
Simulation

We now compute a single realization based on the model we fit in the kriging example.

The input is:

```
USE SOIL
LET CARBON = L10(CARBON)
SPATIAL
MODEL CARBON = LON LAT / SILL=.05,NUGGET=.05,AHMAX=10,
                                        AHMIN=10
SIMULATE / GRAPH=CONTOUR
```

The output is:



The results follow the same pattern found in the kriging. Higher carbon levels occur in the northeast and northwest. The time to compute a single simulation is greatly affected by the number of grid nodes specified in the GRID command. Grids larger that 10 cuts per variate, particularly for larger data sets, can increase the memory and time requirements substantially.

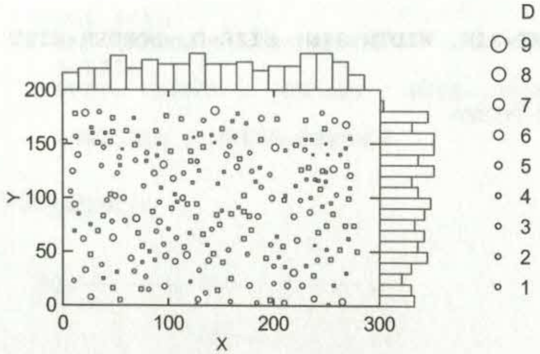
Example 3
Point Statistics

The data for this example are from Kooijman (1979), reprinted in Upton and Fingleton (1990). They consist of the locations of beadlet anemones (Actinia equina) on the surface of a boulder at Quiberon Island, off the Brittany coast, in May, 1976. We have added bordering histograms to the scatterplot shown in Figure 1.26 of Upton and Fingleton. The size of the points is proportional to the measured diameter of the anemones (D).

The input is:

```
USE KOOIJMAN
PLOT Y*X / HEIGHT=2IN, WIDTH=3IN, SIZE=D, BORDER=HIST
```

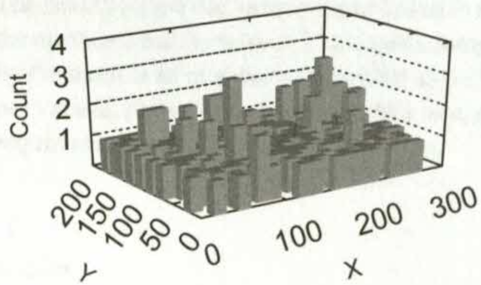
The Output is:

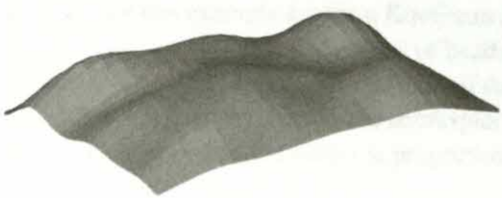


The bordered histograms reveal that the distribution of the anemones is fairly uniform in both marginal directions.

We can get an elevated view of the distribution by computing a 3-D histogram of the anemone locations. A 3-D density kernel provides a smooth estimate of the density. It is available in the SYSTAT graphing module:

```
DEN .*Y*X / HEIGHT=2IN, WIDTH=3IN, ALT=2IN, AXES=CORNER,
            ZGRID
DEN .*Y*X / KERNEL, HEIGHT=2IN, WIDTH=3IN, AXES=0,
            SCALES=0
```

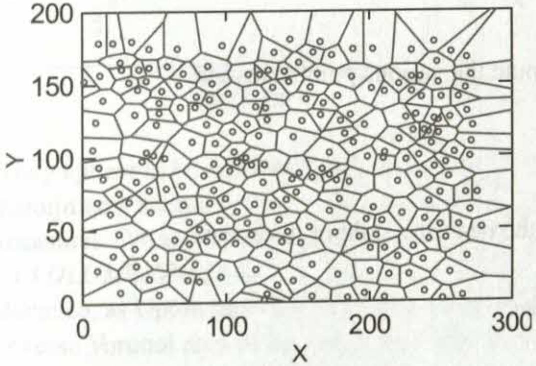


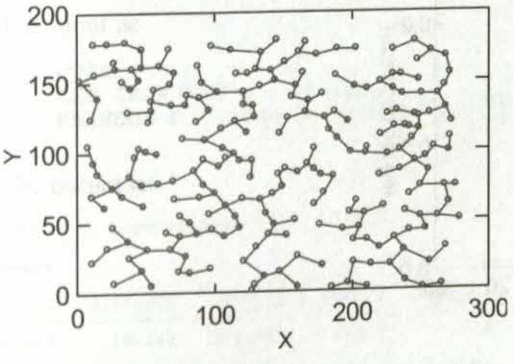


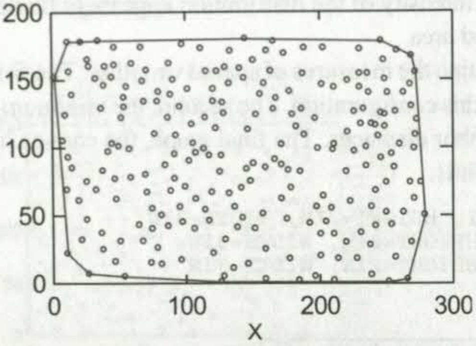
With a few exceptions, the intensity of the distribution appears to be fairly uniformly distributed over the sampled area

Now we proceed to examine the measures of spatial variation. The first graph shows the Voronoi tessellation of this configuration. The second, the minimum spanning tree, highlights the nearest-neighbor distances. The final graph, the convex hull, highlights the outermost bordering points:

```
PLOT Y*X / VORONOI, HEIGHT=2IN, WIDTH=3IN
PLOT Y*X / SPAN, HEIGHT=2IN, WIDTH=3IN
PLOT Y*X / HULL, HEIGHT=2IN, WIDTH=3IN
```



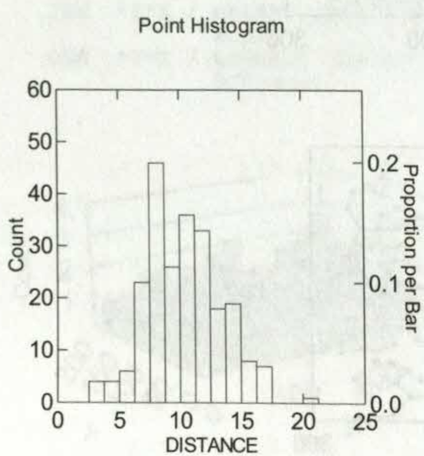




Now we proceed to compute the various point statistics:

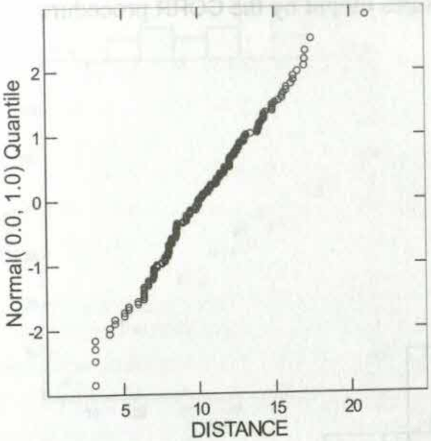
```
SPATIAL
SAVE POINTS
POINT X Y
```

Here is the histogram of the nearest-neighbor distances:



We can do a probability plot of these distances by merging the point measures with the original data:

```
MERGE KOOIJMAN POINTS
PPLOT DISTANCE
```



They appear to be quite normally distributed. Now we examine the relation of Kooijman's measurement of the diameter of the anemones to the other spatial measures. We also construct a new variable called *CROWDING* by taking the inverse of *VOLUME*. The *D* variable is Kooijman's anemone diameter. It is correlated with distance, as Upton and Fingleton point out, but even more strongly related to the inverse Voronoi area of the anemones. This relationship holds even after deleting the four outlying values of *CROWDING*.

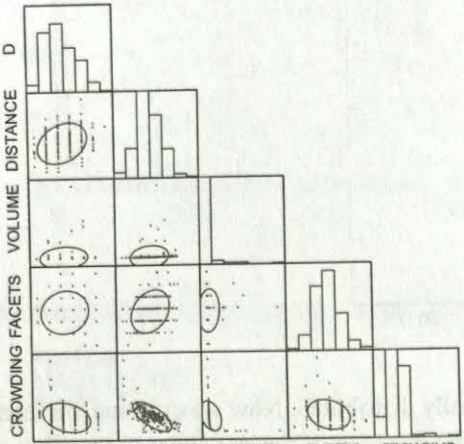
The input is:

```
CORR
LET CROWDING = 1/VOLUME
PEARSON D DISTANCE VOLUME FACETS CROWDING
```

The output is:

```
Number of observations: 217

Means
      D    DISTANCE     VOLUME    FACETS   CROWDING
---------------------------------------------------
  4.263      10.148    954.669     5.829      0.005

Pearson correlation matrix
          |        D   DISTANCE    VOLUME    FACETS   CROWDING
----------+---------------------------------------------------
D         |    1.000
DISTANCE  |    0.249      1.000
VOLUME    |    0.092      0.144     1.000
FACETS    |    0.085      0.285     0.082     1.000
CROWDING  |   -0.291     -0.680    -0.342    -0.237      1.000
```

Here is the SPLOM of these measures output by the CORR procedure:

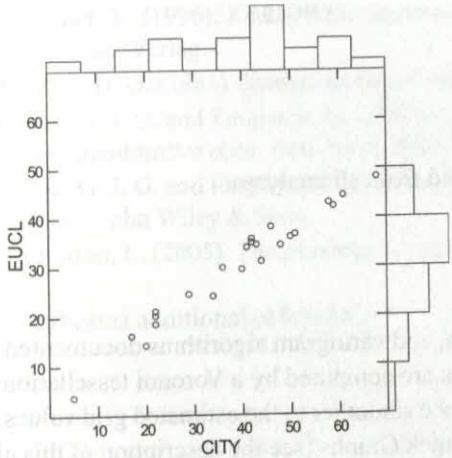


Example 4
Unusual Distances

The transformation and programming capabilities of SYSTAT can be used to compute statistics needed for other spatial analyses. The following example computes Euclidean and city-block distances for the crab data and plots them against each other. The distances are computed from a central point (*px*, *py*) in the field. The city-block distances have particular significance for fiddler crabs in Truro because the encroachment of recent residential and commercial development may force the crabs to follow a rectangular traffic grid to go about their business.

The input is:

```
USE CRABS
LET PY=40
LET PX=40
LET EUCL=SQR((Y-PY)^2+(X-PX)^2)
LET CITY=ABS(Y-PY)+ABS(X-PX)
PLOT EUCL*CITY / BORDER=HIST,XMIN=0,XMAX=70,YMIN=0,YMAX=70
```



Several other useful distance statistics can be calculated directly from coordinate information. The distance between two points on the circumference of a circle given angle coordinates in degrees is:

```
LET CDIST = 2*3.14159*RADIUS*ABS(ANG1-ANG2)/360.
```

The great-circle global distance in statute miles between two points is:

```
REM DEGRAD = RADIANS PER DEGREE
REM AY = NORTH LATITUDE OF POINT A
REM AX = WEST LONGITUDE OF POINT A
REM BY = NORTH LATITUDE OF POINT B
REM BX = WEST LONGITUDE OF POINT B
REM MR = STATUTE MILES PER RADIAN
REM THIS EXAMPLE SETS THE REFERENCE POINT (AX,AY) NEAR CHICAGO
LET DEGRAD=2*3.1415926/360
LET MR=69.09332414/DEGRAD
LET AY=45*DEGRAD
LET AX=-90*DEGRAD
LET BY=LABLAT*DEGRAD
LET BX=LABLON*DEGRAD
LET GCDIST = MR * ACS(SIN(AY)*SIN(BY) ,
                 + COS(AY)*COS(BY)*COS(AX-BX))
```

Computation

Missing Data

Cases with missing data are deleted from all analyses.

Algorithms

SPATIAL uses kriging, simulation, and variogram algorithms documented in Deutsch and Journel (1998). Point statistics are computed by a Voronoi tessellation algorithm. SYSTAT applies the inverse distance smoother to the estimated grid values for kriging and simulation when producing Quick Graphs (see the description of this algorithm in *SYSTAT Graphics*). For sparser grids, this can lead to a high degree of interpolation of the estimated values. To view the actual estimates, save the results into a file and plot them separately without a smoother. You can also specify a large number of grid points (more than 40) to minimize the effects of the inverse smoother.

References

Cressie, N. A. C. (1990). The origins of kriging. *Mathematical Geology*, 22, 239–252.

Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: John Wiley & Sons.

Deutsch, C. V. and Journel, A. G. (1997). *GSLIB: Geostatistical software library and user's guide* (2nd ed.). New York: Oxford University Press.

Haining, R. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge: Cambridge University Press.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining geostatistics*. New York: Academic Press.

Kooijman, S. A. L. M. (1979). The description of point patterns. In R. M. Cormack and J. K. Ord (eds.), *Spatial and Temporal Analysis in Ecology*. Fairland, Md.: International Co-operative Publishing House, pp. 305–332.

Krige, D. G. (1962). Two-dimensional weighted average trend surfaces for ore evaluation. In *Proceedings of the Symposium on Mathematical Statistics and Computer Applications in Ore Valuation*. Johannesburg, 13–38.

Okabe, A., Boots, B., and Sugihara, K. (1992). *Spatial tessellations: Concepts and applications of voronoi diagrams*. New York: John Wiley & Sons.

Pannatier, Y. (1996). *VARIOWIN: Software for spatial data analysis in 2D*. New York: Springer-Verlag.
*Ripley, B. D. (1981). *Spatial statistics*. New York: John Wiley & Sons
Upton, G. J. G. and Fingleton, B. (1985). *Spatial data analysis by example: Point pattern and quantitative data*. New York: John Wiley & Sons.
Upton, G. J. G. and Fingleton, B. (1990). *Spatial data analysis by example* (2 vols.). New York: John Wiley & Sons.
Wilkinson, L. (2005). *The grammar of graphics*. 2nd ed. New York: Springer-Verlag.

(* indicates additional reference.)

[illegible]acier, V. (1996). [illegible] software for spatial data analysis in S. New York: Springer-Verlag.
Ripley, B. D. (1981). Spatial statistics. New York: John Wiley & Sons.
Upton, G. J. G., and Fingleton, B. (1985). Spatial data analysis by example. Vol. 1: Point pattern and quantitative data. New York: John Wiley & Sons.
Upton, G. J. G., and Fingleton, B. (1989). Spatial data analysis by example. Vol. 2. New York: John Wiley & Sons.
Wilkinson, L. (2005). The grammar of graphics. 2nd ed. New York: Springer-Verlag.

(* indicates additional references.)

[illegible]

References

[illegible]

Chapter

13

Survival Analysis

Dan Steinberg, Dale Preston, Doug Clarkson, and Phillip Colla
(Revised by Rajashree Kamath and Amit Kothiyal)

SURVIVAL can be used to explore grouped, right-censored, and interval-censored survival data and to estimate nonparametric, partially parametric, and fully parametric models by maximum likelihood. SURVIVAL can handle disjoint and overlapping interval-censored data and combinations of interval censoring, right censoring, and exact failure times.

The facilities provided in SURVIVAL include the Kaplan-Meier estimator, Turnbull's generalization of the Kaplan-Meier estimator for interval-censored data, Nelson-Aalen cumulative hazard estimator, plots of failure and censoring times, quantile plots for standardized reference distributions, Cox-Snell residual plots (Cox and Snell,1968) for Cox and parametric models, log-rank tests, the proportional hazards (Cox) regression, and the Weibull, lognormal, log-logistic, and exponential regression models. All models can be estimated with or without covariates, either directly or by stepwise regression procedures. Akaike and Bayesian information criteria (AIC, Schwarz's BIC) values are provided for each fitted model. For more information on AIC and Schwarz's BIC in SYSTAT refer, in the chapter on Linear Models, to "Variable Selection" on page 15 in *Statistics II*. The Kaplan-Meier estimator, quantile plots, and Cox regression all permit stratification. The survivor function, hazard function, reliabilities, and quantiles can be generated from parametric models for specific covariate values, and the baseline hazards can be derived from the Cox and stratified Cox models.

The results of most analytic techniques can be saved into SYSTAT files for further manipulation and analysis with other SYSTAT modules.

Statistical Background

SURVIVAL contains a collection of tools for the analysis of survival or reliability data. Typically, the dependent variable is a duration, such as the length of time it takes a woman to conceive after cessation of birth control pills, the survival times of cancer patients on experimental drugs, or the time a motor runs before it fails. The methods have been used to analyze a broad range of topics including unemployment durations, stability of marriages, people's willingness to pay for public goods, and the lengths of pieces of yarn. It could conceivably be used for the modeling of any strictly positive quantity. These topics are also studied under other names—reliability, duration, waiting time, failure time, event history, and transition data analysis are other titles under which survival topics have been discussed. (References are provided below.)

The distinguishing mark of survival analysis, besides the special parametric models typically used, is that the dependent variable can be censored. SURVIVAL allows for two types of such incomplete data: right-censored and interval-censored data. When a case is right-censored, the dependent variable is known to be greater than a specified number, but its true value is not known. When data are subject to interval censoring, failure times may be known only to have occurred within some specified time interval. Left censoring can be handled by SURVIVAL when it coincides with interval censoring with a zero lower bound.

Interval censoring naturally arises in data collected by periodic inspection (Nelson, 1978). For example, a utility company might check gas meters at three-month intervals. A study of the time between meter failures would be conducted on interval-censored data because the exact failure times would never be known. Meters that had failed would only be known to have failed within some three-month interval, and meters that had not failed would be censored.

In general, censoring can occur because a study is ended after a predetermined time period, after a fixed number of failures has occurred, because of periodic inspection, because cases are subject to competing risks (Cox and Oakes, 1984), or for other reasons. A fairly extensive discussion can be found in Lawless (2002). For the methods of this program to be applicable, the censoring scheme should have nothing to do with the future survival of the case. That is, the censoring process cannot be informative. Conditional on having survived to some time t, cases that are censored at that time should be representative of all cases with the same explanatory variables surviving to time t. If the fact that a case is censored provides information about its expected lifetime that distinguishes it from other cases that have not been censored, the assumptions underlying the models estimable with SURVIVAL are violated. For example, censoring will not be independent of future survival if an investigator

removes all persons with good or bad prognoses; results will also be subject to severe bias if patients remove themselves from a study when they feel they are making little progress. (See Cox and Oakes, 1984, or Lagakos, 1979, for further discussion.) For the remainder of this chapter, we assume that the censoring scheme, whatever it may be, is not informative; you should check the conditions under which your data were gathered to ensure that this condition is met prior to analysis.

Graphics

We are going to reproduce (approximately) Figures 2.1 and 2.2 in Parmar and Machin (1995) to give you an idea of how survival measurements differ from other types of data. This should also give you some ideas about using SYSTAT's graphics to produce survival graphs for publication. The first figure shows patients entering a prospective clinical study at different dates, with known survival times indicated by a solid black symbol and censored times by a pale symbol. The input file looks like this:

```
INPUT ENTRY$,DAYS_IN,DAYS_OUT,CENSOR,SURVIVAL
 01/01/91     0     910    0     910
 01/01/91     0     752    1     752
 03/26/91    86    1092    0    1006
 04/26/91   116     452    1     336
 06/23/91   175    1098    1     923
 07/09/91   190     308    1     118
 07/22/91   203     817    0     614
 08/02/91   214     763    1     549
 09/01/91   244    1098    1     854
 10/07/91   280     432    0     152
 12/14/91   348     645    1     297
 12/26/91   360    1001    0     641
ENDINPUT
LET PATIENT=CASE
LET ENTER=DOC(ENTRY$,'MM/DD/YY')
LET EXIT=ENTER + DAYS_OUT - DAYS_IN - 2
DSAVE PMA
```

```
USE PMA
CATEGORY PATIENT,CENSOR
WINDOW 1900
BEGIN
PLOT PATIENT*EXIT / XFORM='MMM. YYYY',
                    XTICK=2,XPIP=12,
                    XMIN=33238,XMAX=34333,
                    TICK=INDENT,SIZE=1.5,
                    YREVERSE,VECTOR=ENTER,PATIENT,
                    HEIGHT=3IN,WIDTH=4IN,SC=2,AX=C,
                    XLAB=' ',YLAB='PATIENT',
                      FILL=CENSOR,LEGEND=NONE
DRAW LINE / FROM=1.4IN,0IN,TO=1.4IN,3IN
DRAW LINE / FROM=3.81IN,0IN,TO=3.8IN,3IN
WRITE 'Patient accrual period' / LOC=.7IN,3.3IN,
                                 HEI=5PT,WID=5PT,
                                    CENTER
WRITE 'Observation only period'/ LOC=2.6IN,3.3IN,
                                 HEI=5PT,WID=5PT,
                                    CENTER
END
```

We have included the raw data so that you can see something about entering and coding time. First of all, if your data source does not have day-of-the-century values (which most spreadsheets use for their time variable), it is easier to import the data as ASCII dates.

Parmar and Machin use British notation in their Table 2.1 for the dates (for example, 26.12.91 instead of 12/26/91). If you want to enter data that way, change the day-of-century conversion in the program from

```
LET ENTER=DOC(ENTRY$,'MM/DD/YY')
```

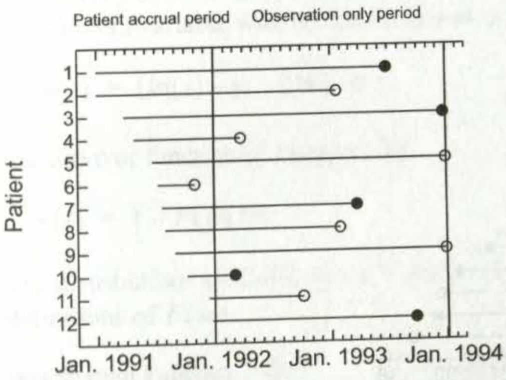
to

```
LET ENTER=DOC(ENTRY$,'DD.MM.YY')
```

Notice how the separator symbol is understood by SYSTAT because you put it in the format string. Any character other than Y, M, D, H, M, or S will do as well. By converting dates to day-of-the-century form, we can now do date arithmetic, calculating the exit times for our graph. We then plot our first graph. Notice, also, how powerful the formatting facility for dates is once we code time as day-of-the-century. We can request the output format for any axis with a simple format string. SYSTAT takes care of choosing round tick-mark values (allowing for leap years and different-length months). Again, if you want another date format, simply change it. For example,

```
XFORM='DD MMM, YYYY'
```

Most of the commands and options are needed to duplicate Parmar and Machin's format. The main idea, however, is that we are seeking a graph that shows how entry and exit times from a study fit on a common time line. Notice, incidentally, that we treat *PATIENT* as a categorical variable, so that each patient is given a tick mark, instead of treating the patient ID's as numbers on a continuous scale. Following is the result:



The second graph changes the time line from calendar time to survival time.

The input is:

```
USE PMA
CATEGORY PATIENT
ORDER PATIENT/SORT=6,10,11,4,8,7,12,2,9,1,5,3
LET ZERO=0
LET SURVIVAL=SURVIVAL/30
PLOT PATIENT*SURVIVAL /,
                 TICK=INDENT,SIZE=1.5,YREVERSE,
                 HEIGHT=3IN,WIDTH=4IN,SC=2,AX=2,
                 XMIN=0,XMAX=36,XTICK=6,XPIP=6,
                 VECTOR=ZERO,PATIENT,FILL=CENSOR,
                 LEGEND=NONE,
                 XLAB='Survival time (months)',
                 YLAB='Patient'
```

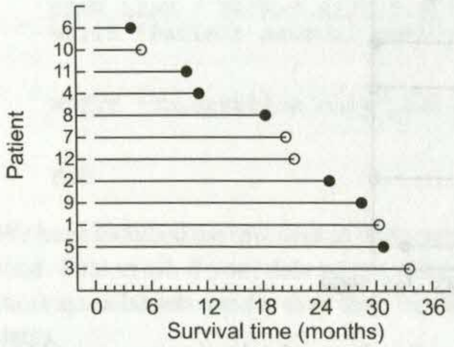
This time, Parmar and Machin order the patients according to survival time, so we use the ORDER command to sort the indices. Since *PATIENT* is categorical, the tick marks will be labeled in that order. We create a *ZERO* value so that we can draw the lines for the dot plot starting at zero survival time.

We used a simple recoding of SURVIVAL to get months:

```
LET SURVIVAL=SURVIVAL/30
```

If we were concerned about accuracy, we could do the time arithmetic exactly with SYSTAT's date functions. (See *SYSTAT*: *Data* for more information.) However, the difference between 30 and 31 days could not be detected in the range of this graph.

The following is the result:



Parametric Modeling

Parametric modeling in SURVIVAL involves the fitting of a fully specified probability model (up to a finite number of unknown parameters) by the method of maximum likelihood. Because the *a priori* commitment to a specific functional form can result in rather poor fits, it is important to explore the fitted model, to examine generalized residuals, and to compare the fitted survivor function to nonparametric and partially parametric models.

The parametric models available in SURVIVAL are based on the exponential, Weibull, lognormal, and log-logistic distributions. Each model can be fit with or without covariates. The exponential and Weibull distributions have two options to allow for the alternative parameterizations discussed below. The Weibull, lognormal, and log-logistic distributions are each specified as two-parameter distributions generalized to include the effects of covariates on survival times. Each is an accelerated life model in which the logarithm of survival time is a linear function of the covariates.

Accelerated Failure Time Distributions

A random variable has an accelerated failure time distribution if the natural logarithm of time can be modeled as

$$\ln(t) = \mu + \beta'\mathbf{z} + \sigma w$$

where μ, β, and σ are parameters to be estimated, z is a vector of covariates, and w is a random variable with the known distribution function $F(w)$. Writing

$$w(t) = (\ln(t) - \mu - \beta'\mathbf{z})/\sigma$$

the survivor function of t is given by

$$s(t) = 1 - F(w(t))$$

The distributions available for accelerated life models in SURVIVAL use the following definitions of $F(w)$:

Distribution Function *F* (w)	**Model**
extreme value: 1-exp[-exp(w)]	EWB, EEXP
logistic: 1 /[1 + exp(-w)]	LGST
standard normal: Φ (w)	LNOR

The Weibull and exponential models can also be estimated in the more familiar proportional hazards parameterization with the WB and EXP commands. The survivor function is now written as

$$s(t) = \exp\left\{-\left(\frac{t}{\pi}\right)^{\delta}\right\}$$

where $\pi = \alpha\ \exp(\mathbf{z}'\beta)$ is proportional to the mean of the distribution and δ is the shape parameter. The exponential distribution is a special case of the Weibull distribution with the shape parameter constrained to 1. In terms of the accelerated life formulation, μ equals $\ln(\alpha)$ and σ equals $1/\delta$. The Weibull distribution is the only distribution that can be equivalently parameterized as either a proportional hazards or an accelerated life model.

Some authors prefer to parameterize the Weibull model in terms of $1/\alpha$ rather than α (for example, Cox and Oakes, 1984). To facilitate comparisons with different texts,

the SURVIVAL output includes several transformations of this parameter. Regardless of the parameterization, the log-likelihood, coefficient estimates, and standard errors for the covariates will be the same. Parameter estimates and standard errors for the location and shape parameters will, however, differ. Choose whatever parameterization is most convenient.

In the output, the fundamental scale (shape) and location parameters are labeled *B(1)* and *B(2)*, respectively. The table below lists their meaning for each of the possible models:

Model	Hazard Shape	B(1)	B(2)
WB	increasing for $\delta > 1$ decreasing for $\delta < 1$	shape δ	scale α proportional to mean time
EWB	increasing for $\sigma > 1$ decreasing for $\sigma < 1$	scale σ	location μ proportional to log mean time
EXP	constant	shape $\delta = 1$	scale α equal to mean time if no covariates
EEXP	constant	scale $\sigma = 1$	location μ equal to log mean time if no covariates
LNOR	non-monotonic	scale σ	location μ
LGST	decreasing for $\sigma > 1$ single-peaked for $\sigma < 1$	scale σ	location μ

Choosing a Parametric Form

Quantile plots of the unadjusted data can be useful in assessing the suitability of a functional form when we are interested in the unconditional distribution of the failure times. When the unconditional distribution may differ substantially from the conditional distribution (conditioning on covariates), the PPLOT output may not be helpful in deciding on a model with covariates. The Cox-Snell residual plot can be used to assess goodness of fit of Cox and parametric models (with or without covariates). You can also examine the quantile Quick Graph plot produced automatically after fitting parametric models.

Selection of a parametric form can also be guided by thinking about the shape of the hazard. This is the approach taken by Barlow and Proschan (1965) and Allison (1984), among others. Since the probability models available in SURVIVAL have sharply different implications for the hazard, any strong prior notions about the hazard time profile can rule out certain models.

The table above lists hazard shapes that are possible for each of the failure models. For example, the exponential model implies a hazard that is constant over time. This means that given a set of covariates, the conditional probability of failure does not depend on the length of the survival time and exhibits duration independence. In contrast, the Weibull model will imply either an increasing or a decreasing hazard, depending on the value of the shape parameter. For example, for much mechanical equipment, the conditional probability of failure is an increasing function of its age (survival time), and a Weibull model is often appropriate.

The remaining two models are a little more complex. The lognormal model yields a nonmonotonic hazard rising to a peak and then declining. If the scale parameter is large, however, the hazard will look like an increasing function over any range of outcomes with appreciable probability. Finally, the log-logistic hazard will be decreasing if the scale parameter is greater than 1; otherwise, it will be non-monotonic with a single maximum (Cox and Oakes, 1984).

It is important to be aware of the potential effect of unobserved heterogeneity on the estimated hazard function. In general, when cases differ along unmeasured dimensions that are relevant to the hazard, the estimated hazard will tend to exhibit a more negative duration dependence than would be obtained with a correctly specified model. For example, Cox and Oakes (1984) point out that if every case has an exponential hazard with a mean parameter distributed as a gamma random variable, the population hazard (a compound exponential) will follow a Pareto distribution with negative duration dependence. This topic has been discussed briefly in the biostatistical literature (Vaupel et al., 1979; Hougaard, 1984; Manton et al., 1986) and has received considerable attention in the econometric literature. See, for example, the *Journal of Econometrics*, Vol. 28 (1985), which is devoted to duration analysis.

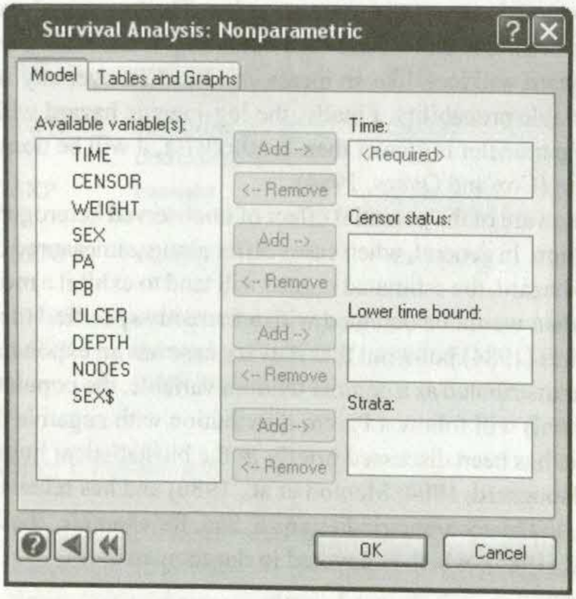
Survival Analysis in SYSTAT

Survival analysis is performed by specifying a model and estimating it. This is true for both parametric models, such as the Weibull, and for nonparametric models, such as Cox regression and Kaplan-Meier curves. For all models, including Kaplan-Meier, Nelson-Aalen and others without covariates, specifying a model may simply be a way of naming survival, censoring, and strata variables. Post hoc analyses, such as plotting survivor functions, computing life tables from a model, and requesting quantiles, are also available.

Survival Analysis: Nonparametric Dialog Box

To open the Survival Analysis: Nonparametric dialog box, from the menus choose:

Advanced
  Survival Analysis
    Nonparametric...



Time. Select the survival variable. This variable is usually a measurement of time, such as the survival duration of a cancer patient or the length of a spell of unemployment, but it could be a weight, a trip length, or any other variable for which negative or zero values would be meaningless.

Censor status. Select the censoring indicator variable. This variable is usually an indicator variable coded as 1 for durations that are complete (uncensored) and 0 for durations that are incomplete (censored). The censoring variable is sometimes called an event variable because it indicates whether or not an event, such as a birth or death, was observed. Survival analysis allows for but does not require censored data; if your observations are all current, each case would have the censoring variable equal to 1.

Lower time bound. Select the lower-bound variable. This variable is used for interval censoring; it need not appear in the data set if the data are subject to right censoring and exact failures alone.

The coding of the survival variable, the censoring variable, and the lower-bound variable depends on whether the data are interval-censored or not. Use the following coding scheme:

Case status	Survival variable	Censoring variable	Lower-bound variable
Exact failure	failure time	1	
Right censored	censoring time	0	
Interval censored	upper bound	–1	lower bound

If the lower-bound variable is specified, it should be coded according to the above scheme. Certain internal data changes are made to the lower bound and censoring variables as the data are entered. For exact failures, the lower bound, if it is included, is set equal to the survival variable. For right-censored cases, the lower bound, if it is being input, is set to –1. For interval-censored cases, the lower-bound value should be non-negative and less than or equal to the survival variable value. If SURVIVAL finds an interval-censored observation with the lower bound equal to the survival time, the censoring is changed to an exact failure (the censoring variable is set to 1). These changes are made solely for the convenience of SURVIVAL, and you will see them only if you save the data during the input process.

Strata. If you want to perform stratified analysis, select the stratification (blocking) variable.

Tables and Graphs

You can select from a variety of output tables when you click the Tables and Graphs tab.

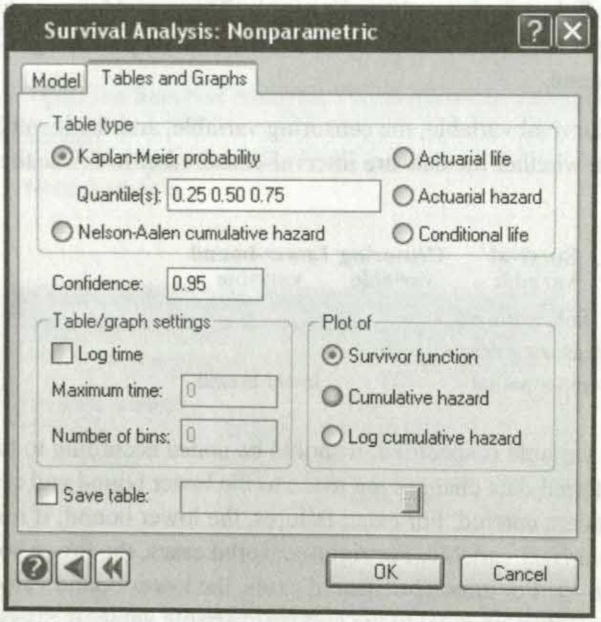


Table type. Select the type of table you would like displayed. The following types are available:

- **Kaplan-Meier probability.** This is a simple nonparametric estimator that produces a table of Kaplan-Meier probabilities, a table of survival quantiles, and a plot of the estimated survivor curve. Optionally, you can specify the percentage points (separated by commas or spaces) at which the quantiles will be reported.
- **Nelson-Aalen cumulative hazard.** This is a nonparametric estimator that produces a table of Nelson-Aalen cumulative hazards and a plot of the estimated cumulative hazard curve.
- **Actuarial life.** Divides the time period of observations into time intervals. Within each interval, the number of failing observations is recorded.
- **Actuarial hazard.** Requests that the hazard function be tabled instead of the standard actuarial survival curve.
- **Conditional life.** Requests that the conditional survival be tabled instead of the standard actuarial survival curve. This table displays the probability of survival given an interval.

Confidence. By default, 95% confidence intervals of the K-M probabilities, the mean survival time, the survival quantiles, and the N-A cumulative hazards are displayed. You can specify a different confidence level.

In addition, you can specify the following options:

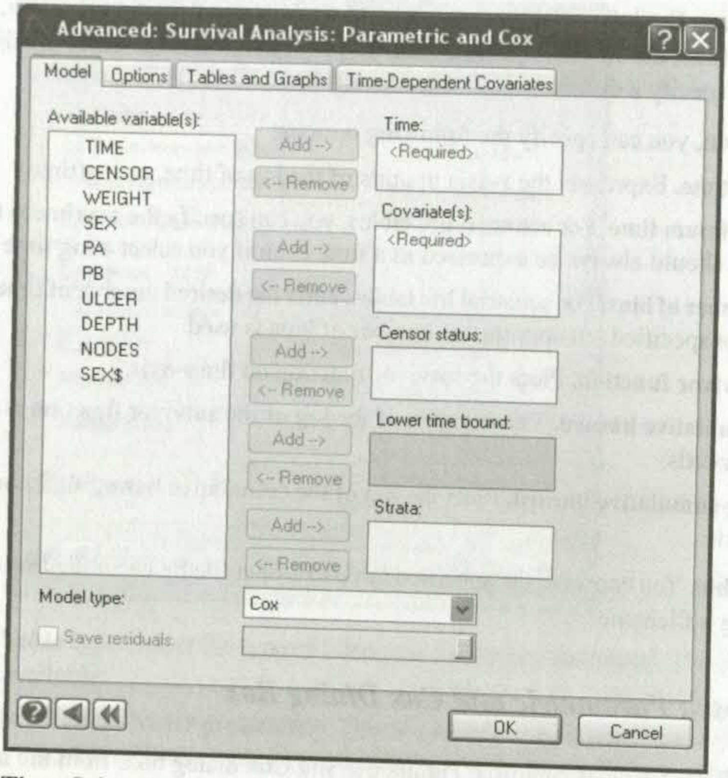
- **Log time.** Expresses the *x*-axis in units of the log of time, or log(time).
- **Maximum time.** For actuarial life tables, you can specify the maximum time limit. This should always be expressed as a time even if you select a log time axis.
- **Number of bins.** For actuarial life tables, enter the desired number of time intervals. If not specified, an appropriate number of bins is used.
- **Survivor function.** Plots the survivor function on the *y*-axis.
- **Cumulative hazard.** The negative of the log of the survivor function is plotted on the *y*-axis.
- **Log-cumulative hazard.** Plots the log of the cumulative hazard function on the *y*-axis.

Save table. You can save the specified table to a data file by checking this option and entering a filename.

Survival Analysis: Parametric and Cox Dialog Box

To open the Survival Analysis: Parametric and Cox dialog box, from the menus choose:

Advanced
  Survival Analysis
    Parametric and Cox...



Time. Select the survival variable. This variable is usually a measurement of time, such as the survival duration of a cancer patient or the length of a spell of unemployment, but it could be a weight, a trip length, or any other variable for which negative or zero values would be meaningless.

Covariate(s). Select covariate variables. Covariates are quantitative predictor variables.

Censor status. Select the censoring indicator variable. This variable is usually an indicator variable coded as 1 for durations that are complete (uncensored) and 0 for durations that are incomplete (censored). The censoring variable is sometimes called an event variable because it indicates whether or not an event, such as a birth or death, was observed. Survival analysis allows for but does not require censored data; if your observations are all current, each case would have the censoring variable equal to 1.

Lower time bound. Select the lower-bound variable. This variable is used for interval censoring; it need not appear in the data set if the data are subject to right censoring and exact failures alone.

The coding of the survival variable, the censoring variable, and the lower-bound variable depends on whether the data are interval-censored or not. Use the following coding scheme:

Case status	Survival variable	Censoring variable	Lower-bound variable
Exact failure	failure time	1	
Right censored	censoring time	0	
Interval censored	upper bound	–1	lower bound

If the lower-bound variable is specified, it should be coded according to the above scheme. Certain internal data changes are made to the lower bound and censoring variables as the data are entered. For exact failures, the lower bound, if it is included, is set equal to the survival variable. For right-censored cases, the lower bound, if it is being input, is set to –1. For interval-censored cases, the lower-bound value should be non-negative and less than or equal to the survival variable value. If SURVIVAL finds an interval-censored observation with the lower bound equal to the survival time, the censoring is changed to an exact failure (the censoring variable is set to 1). These changes are made solely for the convenience of SURVIVAL, and you will see them only if you save the data during the input process.

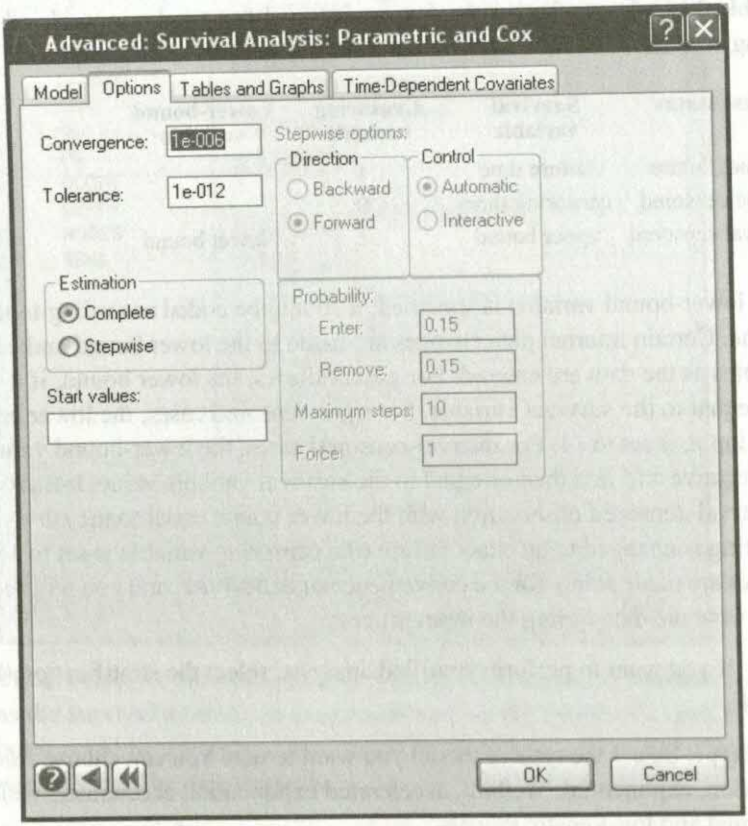
Strata. If you want to perform stratified analysis, select the stratification (blocking) variable.

Model type. Select the type of model you want to use. You can choose from Cox regression, exponential, Weibull, accelerated exponential, accelerated Weibull, lognormal and log-logistic models.

Save Residuals. For AFT models, you can save the regression residuals to a data file.

Options

Click the Options tab in the Survival Analysis: Parametric and Cox dialog box.



The Options tab allows you to specify convergence, a tolerance level, select complete or stepwise estimation, and specify entry and removal criteria.

Convergence. Sets the convergence criterion. This is the largest relative change in any coordinate before iterations terminate.

Tolerance. Prevents the entry of a variable that is highly correlated with independent variables already included in the model. Enter a value between 0 and 1. Typical values are 0.01 or 0.001. The higher the value (closer to 1), the lower the correlation required to exclude a variable.

Estimation. Controls the method used to enter and remove variables from the equation. For complete estimation, in which all independent variables are entered in a single step, you can enter start values. Start values for the computation routines are calculated automatically whenever a model is specified. We suggest that you use these start values unless you have compelling reasons to provide your own or wish to conduct score tests with the Cox model.

For stepwise estimation, in which variables are entered or removed from the model one at a time, the following alternatives are available for stepwise entry and removal:

- **Backward.** Begins with all candidate variables in the model. At each step, SYSTAT removes the variable with the largest Remove value.
- **Forward.** Begins with no variables in the model, and at each step SYSTAT adds the variable with the smallest Enter value.
- **Automatic.** For Backward, at each step SYSTAT automatically removes a variable from your model. For Forward, SYSTAT automatically adds a variable to the model at each step.
- **Interactive.** At each step in the model building process, you select the variable to enter or remove from the model.

Probability. You can also control the criteria used to enter and remove variables from the model:

- **Enter.** Enters a variable into the model if its alpha value is less than the specified value. Enter a value between 0 and 1 (for example, 0.025).
- **Remove.** Removes a variable from the model if its alpha value is greater than the specified value. Enter a value between 0 and 1 (for example, 0.025).

MaxStep. You can define the maximum number of steps that the stepwise estimation should perform.

Force. Forces the first *n* variables listed in your model to remain in the equation, where *n* is the value you enter. The default is 0 for all models except Cox for which it is 1.

Tables and Graphs

You can select from a variety of output tables when you click the Tables and Graphs tab.

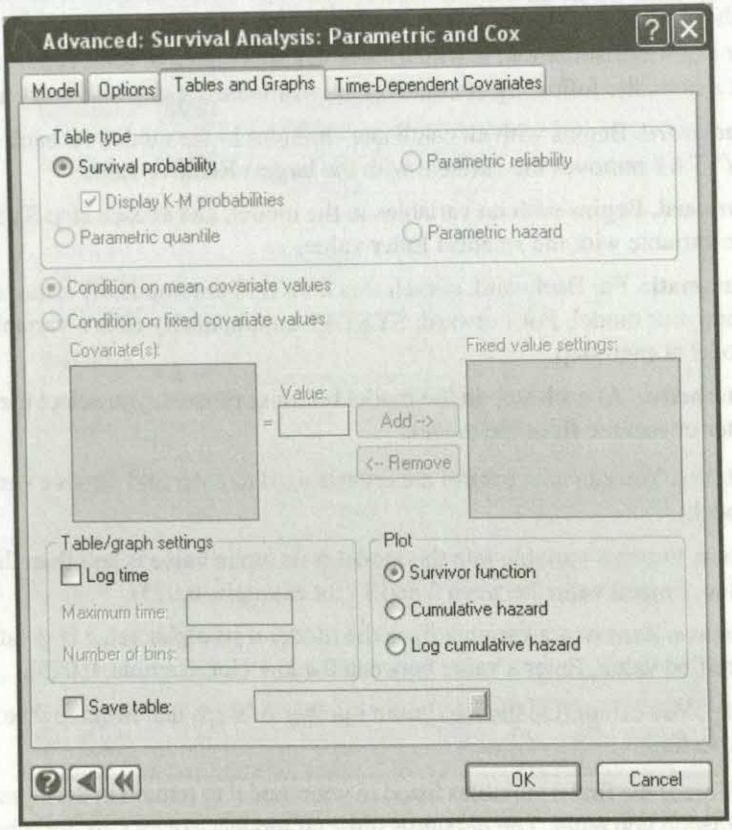


Table type. Select the type of table you would like to be displayed. The following types are available:

- **Survival probabilities.** Produces a table of estimated survival probabilities and a plot of the estimated survivor curve. Optionally, you can request for the corresponding K-M probabilities.
- **Parametric quantiles.** Requests approximate confidence intervals for quantiles and quick graphs based on the last parametric model estimated.

- **Parametric reliability.** Requests reliability confidence intervals and Quick Graphs based on the last parametric model estimated.
- **Parametric hazard.** Requests Quick Graphs and approximate confidence intervals for values of the hazard function at specified times, based on the last parametric model estimated.

Tables, quantiles, hazards, and reliabilities vary as a function of the covariates in the model (if any). SYSTAT offers two methods for dealing with covariates:

- **Condition on mean covariate values.** The survivor curve by default will be evaluated with all covariates set to their means.
- **Condition on fixed covariate values.** You can specify fixed values on the covariates over which tables are produced. Highlight a covariate, enter the fixed value in the Value field, and click Add. The fixed value on the covariate will be displayed in the Fixed value settings list.

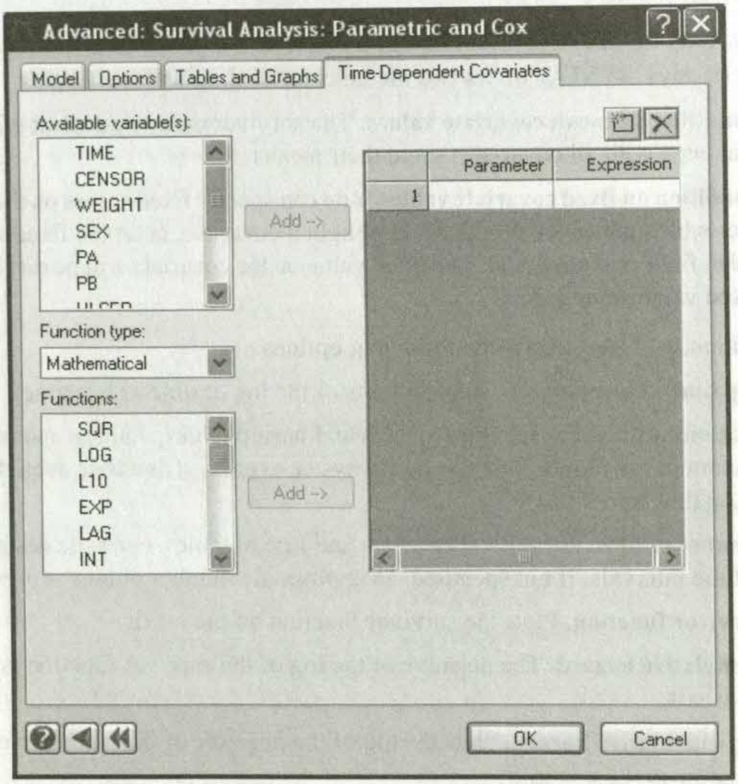
In addition, you can specify the following options:

- **Log time.** Expresses the *x*-axis in units of the log of time or log(time).
- **Maximum time.** For reliability tables and hazard tables, you can specify the maximum time limit. This should always be expressed as a time even if you select a Log time axis.
- **Number of bins.** For reliability tables and hazard tables, enter the desired number of time intervals. If not specified, an appropriate number of bins is used.
- **Survivor function.** Plots the survivor function on the *y*-axis.
- **Cumulative hazard.** The negative of the log of the survivor function is plotted on the *y*-axis.
- **Log-cumulative hazard.** Plots the log of the negative of the cumulative hazard function on the *y*-axis.

Saves table. You can save the specified table to a data file by checking this option and entering a filename.

Time-Dependent Covariates

To specify time-dependent covariates, click the Time-Dependent Covariates tab in the Survival Analysis: Parametric and Cox dialog box.



You can define set names for time-dependent covariates and create, edit, or delete time-dependent covariates.

Parameter. To set a new time-dependent covariate, click the Insert button. Define the covariate under the Expression column. You can use existing variables and choose functions of different types. You may state as many functions of parameters as you want.

You must define a function for each covariate inserted in the Parameter column. When you click OK, SYSTAT will check that each time-dependent covariate has a definition. If a name exists but no variables were assigned to it, the time-dependent covariate is

ignored. To delete a covariate, select the corresponding row and click the Delete button. If you have more than one time-dependent variables, you can rearrange them by selecting a particular row and clicking the Item Up button or Item Down button.

Using Commands

After selecting the data file with USE *filename*, continue with:

```
SURVIVAL
    MODEL timevar = covarlist | tdcovarlist /,
                    CENSOR=var LOWER=var STRATA=var
    FUNPAR tdcovar=expression
(There is one FUNPAR statement for each time-dependent covariate)
    ESTIMATE / method, START=d1,d2,..., TOLERANCE=d,
                    CONVERGE=d
```

Stepwise model fitting is accomplished with the START, STEP, and STOP commands in place of ESTIMATE:

```
START / method, BACKWARD  FORWARD  ENTER=p  REMOVE=p,
                FORCE=n, MAXSTEP=n TOLERANCE=d CONVERGE=d
STEP var or + or - or / AUTO
STOP
```

method is one of:

```
COX   LGST   EXP   EEXP
WB    EWB    LNOR
```

Finally, there are several commands for producing tables and graphs following a model estimation.

```
LTAB / TLOG covar1=d1,covar2=d2,... CHAZ LCHAZ COMP,
QUANTILES=p1,p2,... CONFI=n
NAHAZARD / TLOG CONFI=n LCHAZ
ACT d,n / TLOG  LIFE  CONDITIONAL  HAZARD
QNTL / TLOG
RELIABILITY d,n / TLOG
HAZARD d,n / TLOG
```

Usage Considerations

Types of data. SURVIVAL uses rectangular data and distinguishes three types of data organization, depending on the type of censoring:

- Data are either exact failures or right-censored.
- Interval-censored and right-censored data; intervals do not overlap, and right censoring occurs at the upper boundary of an interval--no exact failures.
- Any other data type, typically, interval-censored data with overlapping intervals, or a mixture of interval-censored and exact failure data.

SURVIVAL automatically classifies the data; the type of data will determine the kinds of analysis you can perform. The fully parametric models can be estimated for any type of data, but the Cox proportional hazards model can be fit only to the first data type, and the K-M estimator is replaced with Turnbull's (1976) generalized K-M estimator for the third data type. When checking for overlapping intervals, SURVIVAL does not consider a shared endpoint to be an overlap.

Categorical variables are used only for stratification. If you have categorical covariates, recode them with the 'Groups and Intervals' functions in SYSTAT before using SURVIVAL.

Print options. PLENGTH LONG adds the 95% confidence intervals and the covariance matrices of parameters to the output.

Quick Graphs. Quick Graphs produced by SURVIVAL include Kaplan-Meier curves, Cox-Snell residual plots and survival functions for Cox and parametric models, probability and quantile plots for parametric models, and reliability and hazard plots for AFT models.

Saving files. Almost every command in SURVIVAL allows you to save selected output to a SYSTAT data file. A SAVE command before ESTIMATE for AFT models saves the residuals. Any command that produces a table or a plot permits a prior SAVE command; this is especially useful if you wish to pursue another type of analysis not presently supported within SURVIVAL.

BY groups. BY group analysis is not allowed in SURVIVAL in the usual sense. However, you may specify a *STRATA* variable, in which case, life tables, quantiles and the Nelson-Aalen estimator will be provided for each group. The only difference is that a single overlaid Quick Graph is produced, and in case of proportional hazards models, the estimated baseline hazard function is common across all strata.

Case frequencies. FREQ variable increases the number of cases by the FREQ variable. It does not use extra memory.

Case weights. WEIGHT is available in SURVIVAL for type 3 data.

The *MELANOMA* file, used in the following examples, is from Hosmer and Lemeshow (2002).

Examples

Example 1
Life Tables: The Kaplan-Meier Estimator

One of the nonparametric estimators available in SURVIVAL is the Kaplan-Meier or the product limit estimator (for example, Kaplan and Meier, 1958, or Lee, 1980). Notice that we estimate a model based solely on time and the censoring structure and then ask for the survival table with the LTAB command. The default survival table produced by LTAB is Kaplan-Meier.

The input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME/CENSOR=CENSOR
ESTIMATE
LTAB/QUANTILES=0.50 0.75 0.90 CONFI=0.99
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records                : 69
Records Kept for Analysis    : 69

Censoring      | Observations
---------------+-------------
Exact Failures |          36
Right Censored |          33

Type 1: Exact Failures and Right Censoring

Overall Time Range:       [72.000 , 7307.000]
Failure Time Range:       [72.000 , 1606.000]

Nonparametric Estimation
Table of Kaplan-Meier Probabilities

All the Data will be Used
```

Number at Risk	Number Failing	Time	K-M Probability	Standard Error	99.0% Confidence Interval Lower	Upper
69.000	1.000	72.000	0.986	0.014	0.825	0.999
68.000	1.000	125.000	0.971	0.020	0.834	0.995
67.000	1.000	127.000	0.957	0.025	0.821	0.990
66.000	1.000	133.000	0.942	0.028	0.805	0.984
65.000	1.000	142.000	0.928	0.031	0.788	0.977
64.000	1.000	151.000	0.913	0.034	0.771	0.969
63.000	1.000	154.000	0.899	0.036	0.753	0.960
62.000	1.000	176.000	0.884	0.039	0.736	0.952
61.000	1.000	184.000	0.870	0.041	0.719	0.943
60.000	1.000	229.000	0.855	0.042	0.702	0.933
59.000	1.000	251.000	0.841	0.044	0.685	0.923
58.000	1.000	256.000	0.826	0.046	0.669	0.913
57.000	1.000	320.000	0.812	0.047	0.652	0.903
56.000	1.000	362.000	0.797	0.048	0.636	0.892
55.000	1.000	391.000	0.783	0.050	0.620	0.882
54.000	1.000	414.000	0.768	0.051	0.605	0.871
53.000	1.000	422.000	0.754	0.052	0.589	0.860
52.000	1.000	434.000	0.739	0.053	0.573	0.848
51.000	1.000	441.000	0.725	0.054	0.558	0.837
49.000	1.000	465.000	0.710	0.055	0.543	0.825
48.000	1.000	471.000	0.695	0.055	0.527	0.813
47.000	1.000	495.000	0.680	0.056	0.512	0.801
45.000	1.000	544.000	0.665	0.057	0.496	0.789
44.000	1.000	584.000	0.650	0.058	0.481	0.776
43.000	1.000	645.000	0.635	0.058	0.466	0.763
42.000	1.000	659.000	0.620	0.059	0.450	0.751
41.000	1.000	749.000	0.605	0.059	0.436	0.738
39.000	1.000	788.000	0.589	0.060	0.420	0.724
37.000	1.000	803.000	0.573	0.060	0.405	0.710
36.000	1.000	812.000	0.557	0.061	0.389	0.696
32.000	1.000	1020.000	0.540	0.061	0.372	0.681
31.000	1.000	1042.000	0.523	0.062	0.355	0.666
28.000	1.000	1151.000	0.504	0.062	0.336	0.650
26.000	1.000	1239.000	0.484	0.063	0.317	0.633
13.000	1.000	1579.000	0.447	0.068	0.270	0.610
12.000	1.000	1606.000	0.410	0.072	0.228	0.584

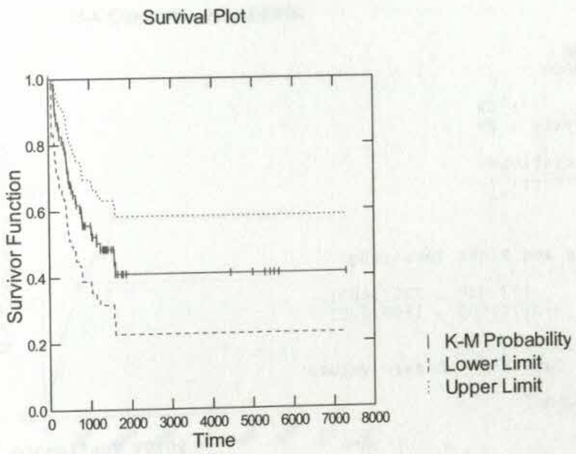
```
Group size                  :    69.000
Number Failing              :    36.000
Product Limit Likelihood    :  -173.084
```

Mean Survival Time

Mean Survival Time	99.0% Confidence Interval Lower	Upper
3371.963	2163.294	4580.631

Survival Quantiles

Probability	Survival Time	99.0% Confidence Interval Lower	Upper
0.500	1239.000	544.000	.
0.750	434.000	176.000	749.000
0.900	154.000	72.000	362.000



The standard error reported for the survivor function is computed using Greenwood's formula (Kalbfleisch and Prentice, 2002). The K-M estimate is a step function with jumps at each exact failure time.

By default, the plots produced by the K-M option are of the survivor function plotted against time. You can also obtain the cumulative hazard plot (time against the negative of the log of the survivor function) or log-cumulative hazard plots (log(–log(survivor))) with the CHAZ and LCHAZ options of the LTAB command.

Example 2
The Nelson-Aalen Estimator

The Nelson-Aalen estimator of the cumulative hazard function is another nonparametric estimator available in SURVIVAL. This estimator can be computed in SYSTAT after model estimation, simply by issuing NAHAZARD instead of LTAB.

The input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME/CENSOR=CENSOR
ESTIMATE
NAHAZARD/CONFI=0.9
```

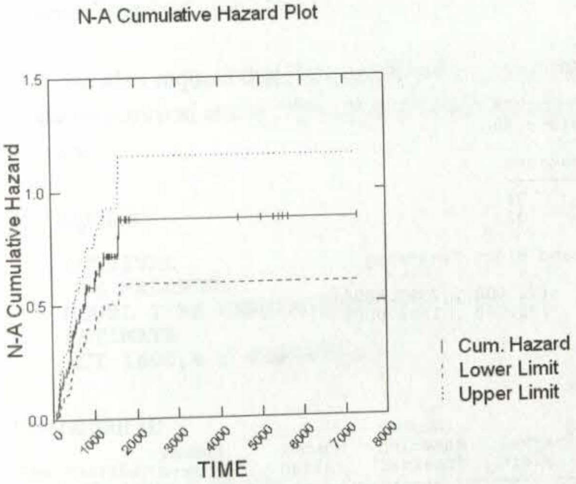
The input above will compute the Nelson-Aalen estimates and their confidence intervals at the observed time points.

The output is:

```
Time Variable    : TIME
Censor Variable : CENSOR

Input Records              : 69
Records Kept for Analysis : 69

Censoring      | Observations
---------------+-------------
Exact Failures |          36
Right Censored |          33

Type 1: Exact Failures and Right Censoring

Overall Time Range:      [72.000 , 7307.000]
Failure Time Range:      [72.000 , 1606.000]

Nonparametric Estimation
Table of Nelson-Aalen Cumulative Hazard Values

All the Data will be Used
```

Number at Risk	Number Failing	Time	N-A Cumulative Hazard	Standard Error	90.0% Confidence Interval Lower	Upper
69.000	1.000	72.000	0.014	0.014	0.000	0.038
68.000	1.000	125.000	0.029	0.021	0.000	0.063
67.000	1.000	127.000	0.044	0.025	0.002	0.086
66.000	1.000	133.000	0.059	0.030	0.011	0.108
65.000	1.000	142.000	0.075	0.033	0.020	0.130
64.000	1.000	151.000	0.090	0.037	0.030	0.151
63.000	1.000	154.000	0.106	0.040	0.040	0.172
62.000	1.000	176.000	0.122	0.043	0.051	0.193
61.000	1.000	184.000	0.139	0.046	0.063	0.215
60.000	1.000	229.000	0.155	0.049	0.074	0.236
59.000	1.000	251.000	0.172	0.052	0.087	0.258
58.000	1.000	256.000	0.190	0.055	0.099	0.280
57.000	1.000	320.000	0.207	0.058	0.112	0.302
56.000	1.000	362.000	0.225	0.060	0.126	0.324
55.000	1.000	391.000	0.243	0.063	0.140	0.347
54.000	1.000	414.000	0.262	0.066	0.154	0.370
53.000	1.000	422.000	0.281	0.068	0.168	0.393
52.000	1.000	434.000	0.300	0.071	0.183	0.416
51.000	1.000	441.000	0.319	0.074	0.198	0.440
49.000	1.000	465.000	0.340	0.076	0.214	0.465
48.000	1.000	471.000	0.361	0.079	0.230	0.491
47.000	1.000	495.000	0.382	0.082	0.247	0.517
45.000	1.000	544.000	0.404	0.085	0.264	0.544
44.000	1.000	584.000	0.427	0.088	0.282	0.571
43.000	1.000	645.000	0.450	0.091	0.301	0.600
42.000	1.000	659.000	0.474	0.094	0.319	0.628
41.000	1.000	749.000	0.498	0.097	0.339	0.658
39.000	1.000	788.000	0.524	0.100	0.359	0.689
37.000	1.000	803.000	0.551	0.104	0.380	0.722
36.000	1.000	812.000	0.579	0.108	0.402	0.756
32.000	1.000	1020.000	0.610	0.112	0.426	0.794
31.000	1.000	1042.000	0.642	0.117	0.450	0.834
28.000	1.000	1151.000	0.678	0.122	0.477	0.879
26.000	1.000	1239.000	0.716	0.128	0.506	0.927
13.000	1.000	1579.000	0.793	0.149	0.548	1.039
12.000	1.000	1606.000	0.877	0.171	0.595	1.158

```
Group size                 :     69.000
Number Failing             :     36.000
Product Limit Likelihood : -173.084
```



Plots of the N-A cumulative hazard can be used to decide on the shape of the hazard function. For instance, the hazard function is constant if the plot is linear, and monotonic if the plot is convex.

Example 3
Actuarial Life Tables

Actuarial life tables divide the time period of observations into time intervals. Within each interval, the number of failing observations is recorded. To see an actuarial table, you must first specify and estimate a model. We already did so in the Life Tables example.

The input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME/CENSOR=CENSOR
ESTIMATE
ACT 1600,4
```

We use the required time parameter to specify the maximum time (1600) and the optional number of intervals (4) to keep the table brief. The default number of intervals is 10.

The output is:

```
Time Variable    : TIME
Censor Variable : CENSOR

Input Records               : 69
Records Kept for Analysis : 69

Censoring       | Observations
----------------+-------------
Exact Failures |           36
Right Censored |           33

Type 1: Exact Failures and Right Censoring

Overall Time Range:      [72.000 , 7307.000]
Failure Time Range:      [72.000 , 1606.000]

Actuarial Life Table

All the Data will be Used
```

Lower Interval Bound	Interval Midpoint	Interval Width	Number Entering Interval	Number Failed	Number Censored
0.000	200.000	400.000	69.000	15.000	0.000
400.000	600.000	400.000	54.000	13.000	4.000
800.000	1000.000	400.000	37.000	5.000	6.000
1200.000	1400.000	400.000	26.000	2.000	12.000

```
1 observed failures are outside the Range of time intervals requested;
they are included in the calculation of the numbers exposed to risk.
```

Hazard Function

You can request that the hazard function be tabled instead of the standard actuarial survival curve.

The input is:

```
ACT 1600,4 / HAZARD
```

The output is:

```
Actuarial Hazard Table

All the Data will be Used
```

Lower Interval Bound	Interval Midpoint	PDF	Standard Error PDF	Hazard Rate	Standard Error Hazard Rate
0.000	200.000	0.001	0.002	0.001	0.000
400.000	600.000	0.000	0.002	0.001	0.000
800.000	1000.000	0.000	0.002	0.000	0.000
1200.000	1400.000	0.000	0.002	0.000	0.000

```
1 observed failures are outside the Range of time intervals requested;
they are included in the calculation of the numbers exposed to risk.
```

Conditional Survival

You can also request that the conditional survival be tabled instead of the standard actuarial survival curve. This table displays the probability of survival given an interval.

The input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME/CENSOR=CENSOR
ESTIMATE
ACT 1600,4 / CONDITIONAL
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records                : 69
Records Kept for Analysis    : 69

Censoring       | Observations
----------------+-------------
Exact Failures  |          36
Right Censored  |          33

Type 1: Exact Failures and Right Censoring

Overall Time Range:       [72.000 , 7307.000]
Failure Time Range:       [72.000 , 1606.000]

Conditional Life Table

All the Data will be Used
```

Interval Midpoint	Number Exposed to Risk	Conditional Probability of Failure Within Interval	Conditional Probability of Failure Beyond Interval	Cumulative Probability of Survival to Beginning of Interval	Standard Error of Cumulative Probability of Survival
200.000	69.000	0.217	0.783	1.000	
600.000	52.000	0.250	0.750	0.783	0.050
1000.000	34.000	0.147	0.853	0.587	0.060
1400.000	20.000	0.100	0.900	0.501	0.062

```
1 observed failures are outside the Range of time intervals requested;
they are included in the calculation of the numbers exposed to risk.
```

Example 4
Stratified Kaplan-Meier Estimation

Nonparametric analysis can be refined further by the use of a stratification variable. There is no fixed upper limit on the number of strata that can be used, and with

moderate-sized data sets, a large number of strata are possible. In the *MELANOMA* data set, the variable *SEX* is coded as 1 for males and 0 for females, respectively. By adding the STRATA option, we will get a single plot with the two estimated survivor curves:

```
SURVIVAL
USE MELANOMA
MODEL TIME / CENSOR=CENSOR, STRATA=SEX
LABEL SEX / 1=''Male'',0=''Female''
PLENGTH LONG
ESTIMATE
LTAB
```

The output is:

```
Time Variable    : TIME
Censor Variable : CENSOR

Input Records                : 69
Records Kept for Analysis : 69

Censoring       | Observations
----------------+-------------
Exact Failures  |           36
Right Censored  |           33

Type 1: Exact Failures and Right Censoring

Overall Time Range:       [72.000 , 7307.000]
Failure Time Range:       [72.000 , 1606.000]

Stratification on SEX specified, 2 levels

Nonparametric Estimation
Table of Kaplan-Meier Probabilities
With stratification on SEX

The following results are for SEX = Female.
```

Number at Risk	Number Failing	Time	K-M Probability	Standard Error	95.0% Confidence Interval Lower	Upper
31.000	1.000	133.000	0.968	0.032	0.792	0.995
30.000	1.000	184.000	0.935	0.044	0.766	0.983
29.000	1.000	251.000	0.903	0.053	0.729	0.968
28.000	1.000	320.000	0.871	0.060	0.692	0.950
27.000	1.000	391.000	0.839	0.066	0.655	0.929
26.000	1.000	414.000	0.806	0.071	0.619	0.908
25.000	1.000	434.000	0.774	0.075	0.584	0.885
23.000	1.000	471.000	0.741	0.079	0.547	0.861
22.000	1.000	544.000	0.707	0.082	0.512	0.836
20.000	1.000	788.000	0.672	0.085	0.475	0.808
19.000	1.000	812.000	0.636	0.088	0.439	0.780
15.000	1.000	1151.000	0.594	0.092	0.394	0.747
13.000	1.000	1239.000	0.548	0.095	0.346	0.711
5.000	1.000	1579.000	0.438	0.124	0.199	0.657
4.000	1.000	1606.000	0.329	0.133	0.103	0.580

```
Group size                 :  31.000
Number Failing             :  15.000
Product Limit Likelihood : -58.200

Mean Survival Time
```

Mean Survival Time	95.0% Confidence Interval Lower	Upper
2395.302	1278.588	3512.017

Survival Quantiles

Probability	Survival Time	95.0% Confidence Interval Lower	Upper
0.250	.	1579.000	.
0.500	1579.000	788.000	.
0.750	471.000	251.000	1151.000

The following results are for SEX = Male.

Number at Risk	Number Failing	Time	K-M Probability	Standard Error	95.0% Confidence Interval Lower	Upper
38.000	1.000	72.000	0.974	0.026	0.828	0.996
37.000	1.000	125.000	0.947	0.036	0.806	0.987
36.000	1.000	127.000	0.921	0.044	0.775	0.974
35.000	1.000	142.000	0.895	0.050	0.743	0.959
34.000	1.000	151.000	0.868	0.055	0.712	0.943
33.000	1.000	154.000	0.842	0.059	0.682	0.926
32.000	1.000	176.000	0.816	0.063	0.652	0.908
31.000	1.000	229.000	0.789	0.066	0.623	0.889
30.000	1.000	256.000	0.763	0.069	0.594	0.869
29.000	1.000	362.000	0.737	0.071	0.566	0.849
28.000	1.000	422.000	0.711	0.074	0.539	0.828
27.000	1.000	441.000	0.684	0.075	0.511	0.807
26.000	1.000	465.000	0.658	0.077	0.485	0.785
25.000	1.000	495.000	0.632	0.078	0.459	0.763
23.000	1.000	584.000	0.604	0.080	0.431	0.739
22.000	1.000	645.000	0.577	0.081	0.405	0.715
21.000	1.000	659.000	0.549	0.081	0.378	0.691
20.000	1.000	749.000	0.522	0.082	0.353	0.666
18.000	1.000	803.000	0.493	0.082	0.325	0.640
16.000	1.000	1020.000	0.462	0.083	0.297	0.612
15.000	1.000	1042.000	0.431	0.083	0.269	0.584

Group size : 38.000
Number Failing : 21.000
Product Limit Likelihood : -89.404

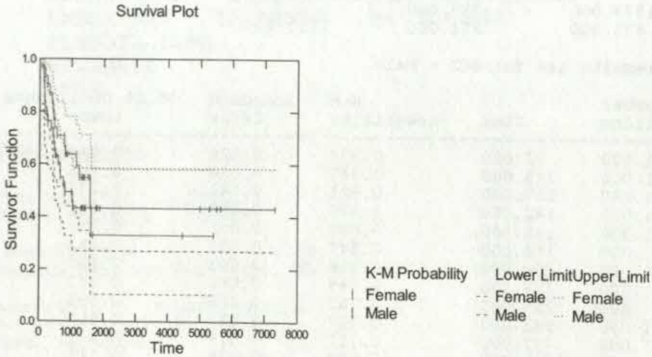
Mean Survival Time

Mean Survival Time	95.0% Confidence Interval Lower	Upper
3404.857	2282.604	4527.110

Survival Quantiles

Probability	Survival Time	95.0% Confidence Interval Lower	Upper
0.250	.	.	.
0.500	803.000	465.000	.
0.750	362.000	142.000	584.000

```
Log-Rank Test, Stratification on SEX, Strata Range 1 to 2

              ¦ Chi-square
              ¦  Statistic
Method        ¦  with 1 df   p-value
--------------+---------------------
Mantel        ¦      0.568     0.451
Breslow-Gehan ¦      1.589     0.207
Tarone-Ware   ¦      1.167     0.280
```



The plot can be used to see if the survivor curves are similar in shape and how far apart they lie. By computing the survivor curves in their log(-log(survivor)) transforms (with the LCHAZ option), you can check for parallelism. Parallel curves, even if the curves themselves are not linear, suggest that the stratification variable acts as a covariate in a proportional hazards model. (See Kalbfleisch and Prentice, 2002, and the Cox regression examples for further discussion of this point.)

If the SAVE command is issued just prior to the LTAB command, a log-cumulative hazard function for each stratum will be saved.

Log-Rank Tests

This output includes three variations of the log-rank test. The first, the Mantel-Haenszel test, is what is conventionally called the log-rank test. The remaining tests are versions of Wilcoxon tests, and they offer different weighting schemes in calculating the difference between observed and expected failures at each failure time in a contingency table analysis. The simple log-rank test uses unit weights so that each failure time has equal weighting. The Breslow-Gehan version weights each failure time by the total number at risk at that time so that earlier times receive greater weight than later times. The Tarone-Ware version weights by the square root of the total number at risk, placing less emphasis on later failure times.

Discussions of log-rank tests can be found in Kalbfleisch and Prentice (2002), Lawless (2002), Miller (1981), and Cox and Oakes (1984). The tests themselves were introduced by Mantel and Haenszel (1959), Gehan (1965), Breslow (1970), and Tarone and Ware (1977). If there are no tied failures, the simple log-rank test is equivalent to a score test of the proportional hazards model containing a dummy variable for each stratum.

Example 5
Turnbull Estimation: K-M for Interval-Censored Data

The Kaplan-Meier estimator, as originally introduced in 1958, is restricted to exact failure and right-censored data or, in other words, type 1 data.

For data of type 2, in which there are disjoint intervals and possibly right censoring, the K-M estimator is extended so that the original definition still applies. Now the number of failures in each interval is considered, and right censoring is assumed to have occurred immediately after the upper boundary of the appropriate interval.

For data of type 3, the generalization of the Kaplan-Meier estimator requires a major departure from the original definition. A version of this generalized K-M estimator was first suggested by Peto (1973) and was further developed by Turnbull (1976). As type 3 data have overlapping interval-censored data and may have exact failures and right censoring as well, the first task is to determine the intervals over which the survivor function is estimated to decrease. Because this estimator is not discussed in the standard texts, we provide a brief exposition of the method here.

When data are of type 3, every case is considered to have left and right time boundaries ($L\{i\}$, $R\{i\}$) defining its interval of censoring or failure. Cases with exact failures have $L\{i\} = R\{i\}$; a right-censored observation will have $R\{i\}$ equal to infinity; and an interval-censored failure will have $L\{i\} < R\{i\}$. The Peto-Turnbull generalization begins by identifying a unique set of disjoint time intervals for which failure probabilities will be estimated. These intervals are constructed by selecting lower boundaries from the left boundaries and upper boundaries from the right boundaries, such that these new intervals do not contain any observed $L\{i\}$ or $R\{i\}$ except at the boundaries.

For example, consider the type 3 data set *TYPE3A*:

LTIME	TIME	WEIGHT	CENSOR
1.0	2.0	4	−1
1.0	2.0	5	−1
1.9	3.0	5	−1
4.0	5.1	3	−1
4.0	4.2	8	−1
5.0	6.0	10	−1
7.0	8.0	6	−1
7.0	9.0	4	−1

There are seven different observed intervals in the data, and the following four intervals are generated by the Turnbull estimator:

lower (q)	upper (p)
1.9	2.0
4.0	4.2
5.0	5.1
7.0	8.0

The lower and upper boundaries are referred to as q's and p's, respectively, by both Peto (1973) and Turnbull (1976). We will explain the determination of the first interval. Cases 1 through 3 overlap each other somewhere on the interval (1.0, 3.0) with distinct left boundaries being 1.0 and 1.9 and distinct right boundaries being 2.0 and 3.0. The interval (1.9, 2.0) is the only interval constructible out of these boundaries that does not itself contain another endpoint. For example, (1.0, 2.0) contains the left endpoint 1.9. The constructed intervals are of minimal size and involve a maximal overlap of cases spanning the interval.

A similar method is used to generate the remaining intervals. Intuitively, the goal is to determine where in the interval does the probability of failure lie. Given that a failure occurs between 1.9 and 3.0, and also that a failure occurs between 1.0 and 2.0, our attempt to assign all the probability to the smallest possible interval leads to the choice of the subinterval (1.9, 2.0).

Turnbull shows that a maximum likelihood nonparametric cumulative distribution function (CDF) can assign probability only to these intervals. Further, for a given set of probability assignments, the likelihood is independent of the behavior of the CDF within the interval, meaning that the CDF may be entirely arbitrary within the interval (Wang, 1987).

The second stage of the generalized Kaplan-Meier estimator computation is to assign probability to each $(q\{i\}, p\{i\})$ interval, which will define the CDF that maximizes the likelihood of the data. The solution vector of probabilities s is obtained by the EM algorithm of Dempster et al. (1977). Specifically, the observed frequency distribution of the data should be equal to the expected frequency, given s.

The input is:

```
SURVIVAL
USE TYPE3A
MODEL TIME / CENSOR=CENSOR, LOWER=LTIME
WEIGHT WEIGHT
ESTIMATE
LTAB
```

The output is:

```
Case frequencies determined by value of variable WEIGHT

Time Variable             : TIME
Censor Variable           : CENSOR
Lower Time Bound Variable : LTIME

Sorting was found to be required on the following special variables
TIME

Input Records              : 8
Records Kept for Analysis  : 8

Censoring          | Observations
-------------------+-------------
Exact Failures     |            0
Right Censored     |            0
Interval Censored  |            8

Type 3: General Censoring (Left Censoring and/or Nondistinct Intervals)

Overall Time Range:        [1.000 , 9.000]
Failure Time Range:        [1.000 , 9.000]

Turnbull K-M estimation

All the Data will be Used

Iteration   Log-Likelihood
--------------------------
0            -60.304
1            -59.757
2            -59.757

Convergence achieved in 2 iterations
Final convergence criterion:    0.000
Final Log-Likelihood          : -59.757

                              Turnbull K-M
Lower Time    Upper Time      Probability     Density Change
------------------------------------------------------------
                                                   0.311
     1.900         2.000            0.689          0.207
     4.000         4.200            0.481          0.259
     5.000         5.100            0.222          0.222
     7.000         8.000            0.000
```

The EM algorithm is frequently slow to converge, but it has the advantage of increasing the likelihood on each iteration. For a theoretical discussion of EM convergence, see Wu (1983).

Example 6
Cox Regression

Proportional hazards regression (Cox, 1972) is a hybrid model--partly nonparametric, in that it allows for an arbitrary survivor function like the Kaplan-Meier estimator, and partly parametric, in that covariates are assumed to induce proportional shifts of the arbitrary hazard function. The Kaplan-Meier (product limit) estimator is equivalent to the Cox model without covariates. In SURVIVAL, Cox models are allowed only for type 1 data. The proportional hazards model is assumed to take the form

$$h(t, z) = b(t)f(z.\beta)$$

where $b(t)$ is the nonparametric baseline hazard, and $f(z.\beta)$ is a parametric shift function of the covariate vector z and the parameter vector β. Typically, $f(z.\beta)$ is specified as $\exp(z'\beta)$, where $z'\beta$ is an inner vector product, and this is the form used in SURVIVAL.

SURVIVAL reports maximum likelihood estimates for β and allows access to $h(t,z)$ and $b(t)$ via the LTAB command. Models are specified with the MODEL command and a list of covariates.

Following is an example, the input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME = ULCER,DEPTH,NODES / CENSOR=CENSOR
ESTIMATE / COX
```

The input above will fit the proportional hazards model with three covariates.

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records              : 69
Records Kept for Analysis : 69

Censoring      | Observations
---------------+-------------
Exact Failures |          36
Right Censored |          33
```

```
Covariate Means

ULCER  | 1.507
DEPTH  | 2.562
NODES  | 3.246

Type 1: Exact Failures and Right Censoring

Overall Time Range:       [72.000 , 7307.000]
Failure Time Range:       [72.000 , 1606.000]

Cox Proportional Hazards Estimation

Convergence : 0.000
Tolerance   : 0.000

Iteration   Step   Log-Likelihood
---------------------------------
0           0      -137.527
1           0      -136.100
2           0      -127.887
3           0      -127.813
4           0      -127.813

Results after 4 Iterations

Final Convergence Criterion         : 0.000
Maximum Gradient Element            : 0.000
Initial Score Test of Regression    : 37.083 with 3 df
Significance Level (p-value)        : 0.000
Final Log-Likelihood                : -127.813
AIC                                 : 261.625
Schwarz's BIC                       : 266.376
-2*[LL(0)-LL(4')] Test              : 19.429 with 3 df
Significance Level (p-value)        : 0.000

Parameter | Estimate   Standard Error       Z      p-value
----------+------------------------------------------------
ULCER     |   -0.776            0.376  -2.063      0.039
DEPTH     |    0.094            0.050   1.885      0.059
NODES     |    0.131            0.053   2.490      0.013

          |             95.0% Confidence Interval
Parameter | Estimate          Lower          Upper
----------+----------------------------------------
ULCER     |  -0.776          -1.514         -0.039
DEPTH     |   0.094          -0.004          0.192
NODES     |   0.131           0.028          0.235

Covariance Matrix

       | ULCER   DEPTH   NODES
-------+----------------------
ULCER  | 0.142
DEPTH  | 0.006   0.002
NODES  | -0.005  0.000   0.003

Correlation Matrix

       | ULCER   DEPTH   NODES
-------+----------------------
ULCER  | 1.000
DEPTH  | 0.293   1.000
NODES  | -0.255  -0.052  1.000
```

We are provided with a summary of the iteration log. The partial likelihood began at -137.527 when the parameter vector was all 0's and ended at -127.813.

The output also reports the score test of the hypothesis that all three coefficients are equal to their start values (in this case 0); the chi-square statistic is 37.083 with three degrees of freedom and has a p value less than 0.001. This test is analogous to the F-test reported by the SYSTAT module GLM for a linear regression, and is simply a test of the hypothesis that the gradient of the log-likelihood function is 0 when evaluated at the start values of the coefficients.

Since the start values are 0 for the Cox model (unless specifically set otherwise by the user), the statistic yields a test of a standard null hypothesis. (Other null hypotheses could be conveniently tested by using the START option of the ESTIMATE command.) Asymptotically, the score test above is equal to the likelihood-ratio test defined as twice the difference between the final and initial likelihood values. In larger samples, there is typically good agreement between the two tests. In small samples, as in this case, the statistics may be quite different.

When comparing the parameter estimates of a Cox model with those of a fully parametric model such as the Weibull, it is important to note that the coefficients are expected to have opposite signs and will differ by a scale factor. If the data actually follow a Weibull model with coefficients β and shape parameter σ, then the proportional hazards parameter will be $(-\beta / \sigma)$ (Kalbfleisch and Prentice, 2002).

Example 7
Stratified Cox Regression

The proportional hazards assumption implies that groups with different values of the covariates have unchanging relative hazard functions over time. Thus, in a study of male and female survival, the ratio of male to female hazard functions would be assumed constant if sex were a covariate. If we thought the hazard function for males was increasing relative to the hazard function for females over time, we would have a violation of the proportional hazards assumption for the *SEX* variable.

To accommodate such potential assumption violations, an important generalization of the Cox model is allowed in SURVIVAL. This is the use of stratification (sometimes also referred to as *blocking*). Stratification relaxes the assumption of a single underlying baseline hazard for the entire population. Instead, it permits each stratum to have its own baseline hazard, with considerably different stratum-specific time profiles possible. Stratification stops short of estimating a separate model for each group because the coefficients for the covariates are common across all the strata. To

estimate a stratified Cox model, we proceed with the following input. We have added an LTAB command to plot a cumulative hazard life table for the model.

```
SURVIVAL
USE MELANOMA
MODEL TIME = ULCER,DEPTH,NODES /,
             CENSOR=CENSOR,STRATA=SEX
ESTIMATE / COX
LTAB / ULCER=0,DEPTH=0,NODES=0,LCHAZ
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records              : 69
Records Kept for Analysis  : 69

Censoring       | Observations
----------------+-------------
Exact Failures  |          36
Right Censored  |          33

Covariate Means

ULCER | 1.507
DEPTH | 2.562
NODES | 3.246

Type 1: Exact Failures and Right Censoring

Overall Time Range:        [72.000 , 7307.000]
Failure Time Range:        [72.000 , 1606.000]

Stratification on SEX specified, 2 levels

Cox Proportional Hazards Estimation
With stratification on SEX

Iteration  Step   Log-Likelihood
--------------------------------
0          0        -112.564
1          0        -108.343
2          0        -103.570
3          0        -103.533
4          0        -103.533

Results after 4 Iterations

Final Convergence Criterion          : 0.000
Maximum Gradient Element             : 0.000
Initial Score Test of Regression     : 32.533 with 3 df
Significance Level (p-value)         : 0.000
Final Log-Likelihood                 : -103.533
AIC                                  : 213.066
Schwarz's BIC                        : 217.816
-2*[LL(0)-LL(4')] Test               : 18.063 with 3 df
Significance Level (p-value)         : 0.000

                       Standard
Parameter | Estimate     Error         Z    p-value
----------+------------------------------------------
ULCER     |   -0.817     0.385    -2.123      0.034
DEPTH     |    0.083     0.053     1.587      0.112
NODES     |    0.131     0.057     2.289      0.022
```

Parameter	Estimate	95.0% Confidence Interval Lower	Upper
ULCER	-0.817	-1.570	-0.063
DEPTH	0.083	-0.020	0.186
NODES	0.131	0.019	0.243

Covariance Matrix

	ULCER	DEPTH	NODES
ULCER	0.148		
DEPTH	0.006	0.003	
NODES	-0.006	0.000	0.003

Correlation Matrix

	ULCER	DEPTH	NODES
ULCER	1.000		
DEPTH	0.301	1.000	
NODES	-0.287	-0.096	1.000

Life Table for Last Cox Model

The following results are for SEX = 0.
Evaluated at covariate vector:

ULCER : 0.000
DEPTH : 0.000
NODES : 0.000

No Tied Failure Times

Number at Risk	Number Failing	Time	Model Survival Probability	Model Hazard Rate	Log Cumulative Hazard
31.000	1.000	133.000	0.941	0.059	-2.808
30.000	1.000	184.000	0.883	0.062	-2.088
29.000	1.000	251.000	0.826	0.065	-1.655
28.000	1.000	320.000	0.769	0.069	-1.336
27.000	1.000	391.000	0.712	0.074	-1.080
26.000	1.000	414.000	0.658	0.076	-0.870
25.000	1.000	434.000	0.606	0.078	-0.692
23.000	1.000	471.000	0.554	0.086	-0.526
22.000	1.000	544.000	0.501	0.096	-0.368
20.000	1.000	788.000	0.445	0.112	-0.210
19.000	1.000	812.000	0.392	0.117	-0.067
15.000	1.000	1151.000	0.337	0.142	0.085
13.000	1.000	1239.000	0.277	0.177	0.250
5.000	1.000	1579.000	0.159	0.427	0.610
4.000	1.000	1606.000	0.070	0.556	0.976

Group size : 31.000
Number Failing : 15.000

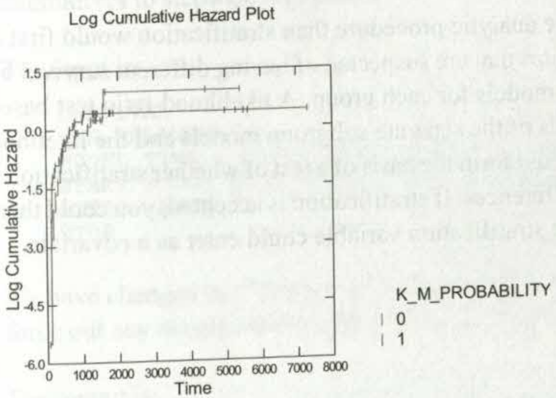
The following results are for SEX = 1.
Evaluated at covariate vector:

ULCER : 0.000
DEPTH : 0.000
NODES : 0.000

No Tied Failure Times

Number at Risk	Number Failing	Time	Model Survival Probability	Model Hazard Rate	Log Cumulative Hazard
38.000	1.000	72.000	0.996	0.004	-5.501

```
 37.000   1.000   125.000      0.953          0.044       -3.023
 36.000   1.000   127.000      0.909          0.046       -2.350
 35.000   1.000   142.000      0.866          0.048       -1.937
 34.000   1.000   151.000      0.823          0.049       -1.637
 33.000   1.000   154.000      0.782          0.050       -1.402
 32.000   1.000   176.000      0.742          0.051       -1.209
 31.000   1.000   229.000      0.703          0.052       -1.044
 30.000   1.000   256.000      0.665          0.055       -0.895
 29.000   1.000   362.000      0.627          0.057       -0.762
 28.000   1.000   422.000      0.590          0.059       -0.639
 27.000   1.000   441.000      0.552          0.063       -0.522
 26.000   1.000   465.000      0.514          0.069       -0.407
 25.000   1.000   495.000      0.476          0.074       -0.298
 23.000   1.000   584.000      0.439          0.077       -0.195
 22.000   1.000   645.000      0.401          0.086       -0.091
 21.000   1.000   659.000      0.361          0.099        0.017
 20.000   1.000   749.000      0.324          0.105        0.121
 18.000   1.000   803.000      0.287          0.113        0.222
 16.000   1.000  1020.000      0.250          0.129        0.326
 15.000   1.000  1042.000      0.215          0.139        0.429

Group size              : 38.000
Number Failing          : 21.000
```




```
Log-Rank Test, Stratification on SEX, Strata Range 1 to 2

                 ! Chi-square
                 !  Statistic
Method           !  with 1 df    p-value
-----------------+-----------------------
Mantel           !    0.568       0.451
Breslow-Gehan    !    1.589       0.207
Tarone-Ware      !    1.167       0.280
```

Stratification allows the survival pattern to vary markedly for cases with different values of the stratification variable while keeping the coefficients governing hazard shifts common across strata. In the above models, allowing *SEX* to be a stratification variable does not alter the coefficients by much.

Comparison of the baseline hazards across the strata allows you to decide whether the stratification variable can be modeled as a covariate. If the log(-log(survivor)) plots

are roughly parallel, the stratification variable is acting to shift the baseline hazard and is correctly considered to be a covariate. If, on the other hand, the curves are quite different in shape, the variable is best left as a stratification variable and should not be included as a covariate. Only one stratification variable can be specified at any given time.

The baseline survivor function derived from the Cox model is produced with the LTAB command followed by zero settings for the covariates, as in our example. By adding the LCHAZ option, we get a baseline hazard for each of the two sexes and a log(-log) plot of the survivor functions against time. As Kalbfleisch and Prentice (2002) point out, this technique can be applied repeatedly, swapping the roles of covariates and stratification variables until you are satisfied with a particular model. With so few data points, it is difficult to draw firm conclusions, but the log(-log(survivor)) curves do look largely parallel. This suggests that *SEX* can appear as a covariate in this model, albeit not a significant one.

A more conservative analytic procedure than stratification would first split the sample into the subgroups that are suspected of having different survival behavior and then estimate separate models for each group. A likelihood-ratio test based on the summed log-likelihoods of the separate subgroup models and the likelihood for the stratified Cox model could form the basis of a test of whether stratification is sufficient to capture the group differences. If stratification is accepted, you could then proceed to investigate whether the stratification variable could enter as a covariate.

Example 8
Stepwise Regression

When there is little theoretical reason to prefer one model specification over another, stepwise methods of covariate selection can be useful, particularly if there is a large number of potential covariates. SURVIVAL allows both forward and backward stepwise covariate selection, with optional forcing of certain covariates into the model and control over the addition and deletion criteria. The stepping can be used with any model (except stratified ones) in SURVIVAL, although the forward selection (STEP/FORWARD) cannot be used with the Cox model unless at least one covariate is forced into the model. In general, we advise you to use backward elimination (STEP/BACKWARD) with all stepwise procedures because it is less likely to miss potentially valuable predictors.

The criterion for adding a variable is based on a Lagrange multiplier test (or Rao's score test) of the hypothesis that the variable has a zero coefficient when added to the

current list of covariates (Peduzzi et al., 1980; Engel, 1984). The signed square root of this chi-square statistic on one degree of freedom is then treated as a normal random variable for significance computation.

For variable deletion, the asymptotic normal statistic based on the ratio of the coefficient to its standard error, as derived from the inverse of the information matrix, is used. If the default ENTER and REMOVE levels are overridden, care should be taken to prevent cycling of variables into and out of the model.

Stepwise model selection in nonlinear contexts is subject to the same criticisms as stepwise linear regression. In particular, conventional hypothesis testing can be misleading, and models will look much better than they really are. For a general discussion of stepwise modeling problems, see Hocking (1983) and additional references cited for General Linear Models. Model selection criteria such as AIC, AIC (corrected) and Schwarz's BIC are also given in the output and could be better alternatives to stepwise regression.

The input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME = ULCER,DEPTH,NODES / CENSOR=CENSOR
START / COX BACK ENTER=0.05 REMOVE=0.05
STEP / AUTO
STOP
```

We have changed the "remove p" value to 0.05 from the default of 0.15 in order to force out any nonsignificant effects from the model .

The output is:

Stepwise Selection of Variables

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records              : 69
Records Kept for Analysis : 69


Censoring       ¦ Observations
----------------+-------------
Exact Failures  ¦           36
Right Censored  ¦           33
```

Covariate Means

```
ULCER ¦ 1.507
DEPTH ¦ 2.562
NODES ¦ 3.246
```

Type 1: Exact Failures and Right Censoring

```
Overall Time Range:       [72.000 , 7307.000]
Failure Time Range:       [72.000 , 1606.000]

Step Number 0

Log-Likelihood  : -127.813

  Information Criteria

  AIC           | 261.625
  Schwarz's BIC | 268.327

  Variables Included

        | Z        p-value
  ------+-----------------
  ULCER | -2.063   0.039
  DEPTH | 1.885    0.059
  NODES | 2.490    0.013

Step Number 1

Log-Likelihood  : -129.259

  Information Criteria

  AIC           | 262.517
  Schwarz's BIC | 266.986
  Variables Included

        | Z        p-value
  ------+-----------------
  ULCER | -2.577   0.010
  NODES | 2.524    0.012

  Variables Excluded

        | Z       p-value
  ------+----------------
  DEPTH | 1.941   0.052

Stepping no longer possible.

Final Model Summary

            |            Standard
  Parameter | Estimate      Error        Z    p-value
  ----------+----------------------------------------
  ULCER     |   -0.926      0.359   -2.577    0.010
  NODES     |    0.136      0.054    2.524    0.012

            |             95.0% Confidence Interval
  Parameter | Estimate          Lower          Upper
  ----------+---------------------------------------
  ULCER     |   -0.926         -1.631         -0.222
  NODES     |    0.136          0.030          0.241

 Covariance Matrix

            | ULCER    NODES
```

```
------+---------------
ULCER ¦ 0.129
NODES ¦ -0.005   0.003
```

Correlation Matrix

```
      ¦ ULCER    NODES
------+---------------
ULCER ¦ 1.000
NODES ¦ -0.280   1.000
```

Example 9
The Weibull Model for Fully Parametric Analysis

This example fits an accelerated life model using the Weibull distribution. When we fit parametric models, we automatically get a Cox-Snell residual plot, and a plot of the log failure times against the quantiles of the chosen distribution.

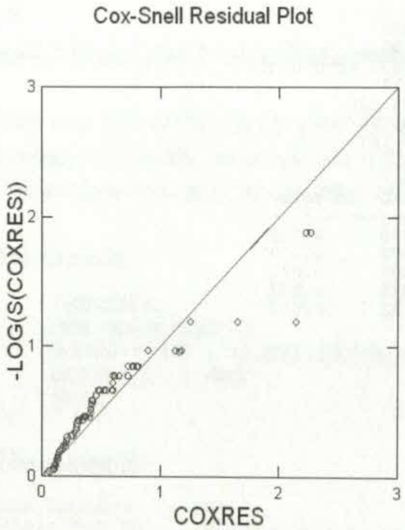
The input is:

```
SURVIVAL
USE MELANOMA
MODEL TIME = ULCER,DEPTH,NODES / CENSOR=CENSOR
ESTIMATE / EWB
QNTL
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records              : 69
Records Kept for Analysis  : 69

Censoring      | Observations
---------------+-------------
Exact Failures |          36
Right Censored |          33

Covariate Means

ULCER | 1.507
DEPTH | 2.562
NODES | 3.246

Type 1: Exact Failures and Right Censoring

Overall Time Range:      [72.000 , 7307.000]
Failure Time Range:      [72.000 , 1606.000]

Weibull Model B(1)--shape, B(2)--scale
Extreme value parameterization

Convergence : 0.000
Tolerance   : 0.000

Iteration   Step   Log-Likelihood   Method
------------------------------------------
0           0      -346.029         BHHH
1           0      -333.961         BHHH
2           0      -325.721         BHHH
3           0      -318.696         BHHH
4           0      -316.158         BHHH
5           0      -312.058         N-R
6           0      -307.552         BHHH
7           0      -306.814         BHHH
8           1      -306.615         N-R
9           0      -306.510         N-R
10          0      -306.508         N-R
11          0      -306.508         N-R

Results after 11 Iterations
```

```
Final Convergence Criterion        : 0.000
Maximum Gradient Element           : 0.000
Initial Score Test of Regression   : 14.738 with 5 df
Significance Level (p-value)       : 0.012
Final Log-Likelihood               : -306.508
AIC                                : 623.016
Schwarz's BIC                      : 634.187

           |              Standard
Parameter  | Estimate     Error          Z     p-value
-----------+------------------------------------------
B(1)       |    1.202     0.161      7.470       0.000
B(2)       |    7.277     0.728      9.990       0.000
ULCER      |    0.776     0.431      1.800       0.072
DEPTH      |   -0.154     0.057     -2.675       0.007
NODES      |   -0.063     0.020     -3.162       0.002

1.0/B(1): 0.832, EXP(B(2)): 1446.887

       | Mean Failure
Vector |        Time      Variance
-------+--------------------------
ZERO   |    1595.592   3716876.337
MEAN   |     900.377   1183539.495

Coefficient of Variation:  1.208

           |              95.0% Confidence Interval
Parameter  | Estimate              Lower        Upper
-----------+-----------------------------------------
B(1)       |    1.202              0.886        1.517
B(2)       |    7.277              5.849        8.705
ULCER      |    0.776             -0.069        1.622
DEPTH      |   -0.154             -0.266       -0.041
NODES      |   -0.063             -0.102       -0.024

Covariance Matrix

       |  B(1)      B(2)      ULCER    DEPTH    NODES
-------+-------------------------------------------
B(1)   |  0.026
B(2)   |  0.003     0.531
ULCER  |  0.007    -0.288     0.186
DEPTH  | -0.001    -0.021     0.007    0.003
NODES  |  0.000    -0.003     0.001    0.000    0.000

Correlation Matrix

       |  B(1)      B(2)      ULCER    DEPTH    NODES
-------+-------------------------------------------
B(1)   |  1.000
B(2)   |  0.024     1.000
ULCER  |  0.108    -0.915     1.000
DEPTH  | -0.132    -0.511     0.291    1.000
NODES  | -0.077    -0.199     0.079    0.020    1.000
```



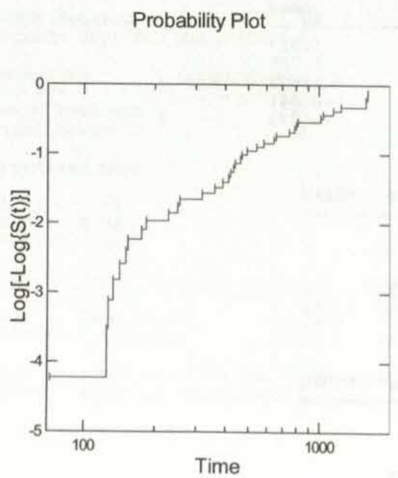


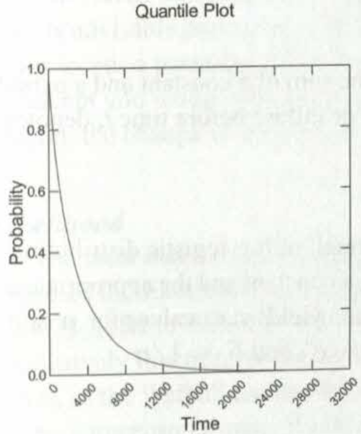
Table of Estimated Quantiles for Last Accelerated Weibull Model

Covariate Vector

ULCER : 1.507
DEPTH : 2.562
NODES : 3.246

Survival Probability	Estimated Time	95.0% Confidence Interval Lower	Upper	Log of Estimated Time	Standard Error of Log Time
0.999	0.637	0.079	5.166	-0.451	1.068
0.995	4.418	0.895	21.825	1.486	0.815
0.990	10.193	2.549	40.769	2.322	0.707

0.975	30.935	10.186	93.952	3.432	0.567
0.950	72.263	29.169	179.023	4.280	0.463
0.900	171.618	84.262	349.534	5.145	0.363
0.750	573.787	353.087	932.437	6.352	0.248
0.667	866.645	560.840	1339.193	6.765	0.222
0.500	1650.688	1101.241	2474.271	7.409	0.207
0.333	2870.859	1861.913	4426.540	7.962	0.221
0.250	3796.547	2386.677	6039.263	8.242	0.237
0.100	6985.190	3989.200	12231.245	8.852	0.286
0.050	9583.149	5152.747	17822.869	9.168	0.317
0.025	12306.215	6287.225	24087.403	9.418	0.343
0.010	16065.792	7752.889	33292.060	9.684	0.372
0.005	19013.916	8840.918	40892.701	9.853	0.391
0.001	26151.527	11313.122	60452.137	10.172	0.428



In the output, the fundamental scale (shape) and location (scale) parameters are labeled *B(1)* and *B(2)*, respectively. Also, notice that SURVIVAL selected the BHHH method in early iterations to ensure a positive definite information matrix. It then switched to the conventional Newton-Raphson.

The Cox-Snell residual plot is used to assess the goodness of fit of a Cox or a parametric model. The Cox-Snell residual method computes the Cox-Snell residuals which are nothing but the estimated cumulative hazard values. If the fitted model is appropriate, a plot of the Cox-Snell residuals versus the K-M cumulative hazard values based on these residuals should be close to a 45° line. Likewise the probability plot should follow a relatively straight line if the distribution used is appropriate. You should compute several distributions and examine these graphs as diagnostic aids. We also present the quantities of the fitted distribution in a table and Quick Graph.

Computation

Algorithms

Start values for the computation routines are calculated automatically whenever a model is specified. In SURVIVAL, start values are obtained from a linear regression based on an accelerated life model without covariates. The model is:

$$\ln(t) = \mu + \sigma w$$

which specifies the log failure time to be the sum of a constant and a parametric error. We rewrite this in terms of the probability of failure before time t, denoted by p, as

$$\ln(t) = \mu + \sigma F(p)^{-1}$$

where F is the CDF of the Weibull, lognormal, or log-logistic distribution. A linear regression of the observed failure times on a constant and the appropriate transform of the Kaplan-Meier estimate of p for each time yields start values for μ and σ. For the WB form of the Weibull model, we use $a = e^{\mu}$ and $\delta = 1/\sigma$.

Missing Data

SURVIVAL will analyze only cases that have valid data for every special variable and all covariates listed in the MODEL command. If any one of these variables is missing for a case, that record will not be input. Consequently, if you want to analyze data containing missing values for some of the covariates and retain the maximum possible number of cases for each analysis, use the CORR procedure to estimate the missing values via the EM algorithm and save your data with the imputed values.

Parameters

In SURVIVAL, we use the accelerated life parameterization for convenience in computing and interpreting the results. The models behave well and converge quickly, and the notion of a covariate accelerating life is intuitive. Some other texts and programs prefer a different parameterization, most typically for the Weibull and exponential models. To facilitate comparisons, SURVIVAL output prints transformations of the scale and location parameters that will match other

parameterizations, and the optional WB and EXP commands use a proportional hazards parameterization. If you observe a difference in the scale and location parameters but identical covariate coefficients (or identical except for sign), you have come across a difference in parameterization. This is no cause for concern; from a mathematical point of view, the sets of results are identical except for a transformation of some parameters.

Centering

In SURVIVAL, the default is to input data without centering. If you do opt to center, and this is advisable particularly for estimation of the WB model, you will discover that your location parameter (*B(2)*) may change. This is analogous to the change in the intercept you would see in a multiple regression if you centered some of your data. Again, the change is of no consequence.

Log-Likelihood

The most common discrepancy between SURVIVAL and textbook results is in the reported log-likelihood at convergence. Some authors such as Kalbfleisch and Prentice (2002) prefer to eliminate any terms in the log-likelihood that are constants or exclusively functions of the data (that is, not functions of the unknown parameters). Thus, in the Weibull model, they drop an $\ln(t)$ term from the likelihood contribution of each uncensored case. While this does not in any way affect the maximum likelihood solutions for the parameter vector, it does result in a log-likelihood much smaller than that reported by SURVIVAL. For example, Kalbfleisch and Prentice (2002) report a Weibull model estimated on the data in their Table 1.2 as having a log-likelihood of –22.952; SURVIVAL reports –144.345. The coefficients and standard errors are identical for both normalizations, however. A similar divergence will be noted on the log-logistic and lognormal models. All of the differences are innocuous and are the result of different normalizations; they do not represent any real differences in results.

Iterations

The maximum likelihood procedures in SURVIVAL are iterative. The basic iteration consists of determining the gradient of the log-likelihood with respect to the parameter vector, calculating a parameter change vector, and evaluating the log-likelihood based on the updated parameter vector. If the new log-likelihood is larger than that of the previous iteration, the iteration is considered complete and a new iteration is begun; if

not, a step halving is initiated. SURVIVAL will continue to iterate until convergence has been attained.

A step-halving is required if some metric of the parameter change vector is too large, resulting in a more negative log-likelihood. This change vector is simply cut in half, and the log-likelihood is reevaluated. If this log-likelihood is an improvement, the iteration is considered complete and a new iteration is begun; otherwise, another step halving is done.

During this process, if either the total number of complete iterations or the total number of step-halvings for a single iteration becomes greater than or equal to a fixed (internal) limit, estimation will stop, and a message stating that the iteration limit was encountered will be given followed by the parameter values, log-likelihood, etc., at this point.

The iteration limit will usually be a problem only for models that typically converge slowly, such as WB and EXP. On the other hand, as the parameter estimates approach their final values and the convergence criterion is almost satisfied, SURVIVAL may have difficulty in improving the log-likelihood. Successively smaller steps will be required to get an improved log-likelihood for the iteration, since there is only a little room left for improvement this far along anyway. If iteration i results in a log-likelihood very close to the optimal value, but the overall convergence criterion is not yet satisfied, then many step-halvings are required on iteration $i + 1$ to get an improvement, and the step-halving limit may be encountered. This may not be a problem. If the parameters that are printed out appear not to have met the convergence criterion, they probably are near their optimal values anyway. Intelligent control of the convergence criterion is important here.

Singular Hessian

SURVIVAL will not estimate models that include an exact linear dependency among covariates or that include a constant covariate. For either situation, the Hessian (matrix of second derivatives) is singular, and a message to that effect will be printed in the output. The problem of covariate interdependency is common to all models (parametric and proportional hazards). Stratified proportional hazards models add another level of complexity. If one of the covariates is constant within a stratum, a singular Hessian can result.

Survival Models

We use the notation $F(t)$ to represent the CDF for the continuous non-negative random variable *T*. Within SURVIVAL, we require that all failure times be strictly positive (that is, zero failure times are not permitted). The survivor function is defined as

$$S(t) = 1 - F(t) = Prob(T > t)$$

The density function is

$$f(t) = dF(t)/(dt)$$

and the hazard function is

$$h(t) = -d[\ln S(t)]/dt = f(t)/S(t)$$

Censoring occurs when the value of *t* is not observed completely but is restricted to an interval on the real line. In general, an observation is interval-censored if all we know is that the failure time falls between times t_u and t_i, or

$$t_i < t < t_u$$

The censoring is called *right censoring* when

$$t_i < t < \infty$$

In some contexts, if $t_i = 0$ and t_u is finite, the censoring is called *left censoring*, but we do not distinguish this from general interval censoring in SURVIVAL.

Proportional Hazards Models

Cox's proportional hazards model can be written as

$$h(t, \mathbf{z}, \beta) = h_0(t)\exp(\mathbf{z}'\beta)$$

where $h_0(t)$ is the nonparametric baseline hazard. The survivor function is then

$$S(t, \mathbf{z}, \beta) = S_0(t)^q$$

where $q = \exp(\mathbf{z}'\beta)$. SURVIVAL allows each stratum i to have its own baseline hazard $h_{0i}(t)$.

The Cox model is estimated by maximizing the partial likelihood that does not include the baseline hazard $h_0(t)$. For tied failure times, we use Breslow's generalization of the Cox likelihood. Denoting the ordered failure times for the ith stratum by $t_{(1i)}, \ldots t_{mi}$

$$L = \prod_{i=1}^{I}\prod_{j=1}^{m_i} \exp(s_{(ji)}\,\beta\,) / \left[\sum_{R_{t(ji)}} \exp(z\,\beta\,)^{d(ji)}\right]$$

where m_i is the number of failures in the ith stratum, $d(ji)$ is the number of failures in stratum i at time $t_{(ji)}$, $s_{(ji)}$ is the vector sum of covariate vectors for each of these $d(ji)$ observations, and R_{tji} is the risk set at failure time $t_{(ji)}$. When there are no tied failures, this formula reduces to Cox's original likelihood.

The recovery of the baseline hazard for a stratum follows Prentice and Kalbfleisch (1979). If there are no tied failure times, defining

$$\alpha_j = \left[1 - \frac{\exp(z'\beta)}{\sum_{R_{tji}} \exp(z'\beta)}\right]^{\exp(-z'\beta)}$$

the baseline hazard for covariate vector $\mathbf{z} = \mathbf{0}$ is

$$S(t;0) = \prod_{t_j < t} \alpha_j$$

For tied failure times,

$$\alpha_j = \exp\left(\frac{-d(ji)}{\sum_{R_{tji}} \exp(z'\beta)}\right)$$

ln(–ln(survivor)) Plots and Quantile Plots

We can write the log(–log(survivor)) equation as

$$\ln(-\ln(S(t, \mathbf{z}, \beta))) = \ln(-\ln(S_0(t))) + \mathbf{z}'\beta$$

which shows that for different values of the covariate vector **z**, the curve is simply shifted by an additive constant $\mathbf{z}'\beta$. Although the baseline curve $\ln(-\ln(S_0(t)))$ need not be linear, the curves for different strata satisfying the proportional hazards assumption will be parallel.

For the Weibull model, the baseline hazard can be written as

$$S(t) = e^{-t^{\delta}}$$

so that

$$\ln(-\ln(S(t))) = \delta \ln(t)$$

which will plot as a straight line against $\log(t)$.

Convergence and Score Tests

The convergence criterion is based on the relative increase in the likelihood between iterations. If

$$[L^{(i)} - L^{(i-1)}]/L^{(i)} < converge$$

convergence is achieved, where $L^{(i)}$ is the value of the log-likelihood at the ith iteration.

The relative change in the log-likelihood is also used to decide whether first derivatives or the Newton-Raphson method is used in the search algorithm. By default, if the relative increase exceeds the user-defined threshold, only first derivatives are calculated, and the sum of the outer products of the gradient vector are used as an approximation to the matrix of second derivatives (Berndt et al., 1974); below the threshold, the Newton-Raphson method is used.

The score test (Rao, 1973; Engel, 1984) is a Lagrange multiplier (LM) test of the hypothesis that the entire parameter vector of the Cox model is **0**. The statistic is computed as

$$S = \mathbf{U}(\delta)'\mathbf{I}(\delta)^{-1}\mathbf{U}(\delta)$$

where $\mathbf{U}(\delta)$ is the score (gradient) vector evaluated at parameter vector δ, and $\mathrm{I}(\delta)$ is an estimate of the information matrix also evaluated at δ. Under the null hypothesis that $\beta = \delta$, S is asymptotically distributed as a chi-square variate with degrees of freedom equal to the number of elements in β. In SURVIVAL, the score test is computed for δ equal to the start values for β. Ordinarily, these are 0 for the Cox model, but they may be overridden with the START option.

Stepwise Regression

The stepwise algorithm follows the suggestion of Peduzzi et al. (1980) and, if unrestricted, begins with a test for downward stepping. The criterion for deletion of a variable is based on the *t* statistic or, more correctly, the asymptotic normal statistic, computed as the ratio of the coefficient to its estimated standard error.

A step up is based on a score test of the hypothesis that a potential covariate not currently in the model has a coefficient of 0. If the model currently has p covariates, to test for the addition of the $(p + 1)$ th covariate, we need to evaluate the information matrix **I** and the score vector **U** under the null hypothesis. Writing β_0 for the current parameter vector obtained from maximizing the log-likelihood for p parameters, and partitioning the score vector $\mathbf{U} = (\mathbf{U}_1,\mathbf{U}_2)$, the score statistic is

$$\mathbf{U}(\beta_0,\mathbf{0})'\mathbf{I}(\beta_0,\mathbf{0})^{-1}\mathbf{U}(\beta_0,\mathbf{0}) = \mathbf{U}_2'\mathbf{I}^{22}\mathbf{U}_2$$

where $\mathbf{I}^{22}$ is the partitioned inverse of **I**. The statistic could be expanded to test for a set of potential covariates jointly but is implemented for a single covariate only in the

current version of SURVIVAL. The resulting scalar is asymptotically a chi-square variate on one degree of freedom, whose square root is treated as a standard normal.

Variances of Quantiles, Hazards, and Reliabilities

The pth quantile of a distribution for the random variable is that value of t for which $F(t) = p$. For the accelerated life model we have

$$\ln(t) = \mu + \beta'\mathbf{z} + \sigma w$$

and for a given p, a point estimate for $\ln(t)$ is obtained from

$$\ln(t) = \mu + \beta'\mathbf{z} + \sigma F^{-1}(w)$$

where F^{-1} is the inverse of the extreme value, normal, or logistic distribution, depending on the model in use. The variance of $\ln(t)$ is derived under the assumption that the estimated parameters are multivariate normal with mean and covariance matrix given by the maximum likelihood solutions. The confidence intervals are computed in terms of $\ln(t)$ and then transformed to the time scale.

Confidence intervals for reliabilities are computed from asymptotic approximations based on a first-order Taylor series expansion in terms of the estimated parameters of the model. This is sometimes called the **delta method** (Rao, 1973). In SURVIVAL, we compute confidence intervals in terms of the log-odds ratio $\ln(p/(1-p))$ because this quantity does not have any range restrictions and is more nearly a linear function of the parameters. The confidence intervals for the log odds are then transformed to the probability scale.

References

* Akaike, H.(1973). Information theory as an extension of the maximum likelihood principle. in B.N. Petrov, and F.Csaki, (eds.) *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp.267-281.

* Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC,* 19, 716-723.

Allison, P. (1984). *Event history analysis*. Beverly Hills, Calif.: Sage Publications.

*Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *Journal of the Royal Statistical Society*, Series B, 42, 322–327.

Barlow, R. E. and Proschan, F. (1965). *Mathematical theory of reliability*. New York: John Wiley & Sons.

Berndt, E. K., Hall, B., Hall, R. E., and Hausman, J. A. (1974). Estimation and inference in non-linear structural models. *Annals of Economic and Social Measurement*, 3, 653–665.

Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika*, 57, 579–594.

* Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89–99.

* Burnham, K.P., and Anderson, D.R. (2002). Model selection and multimodel inference: A practical information-theoretic approach. New York: Springer-Verlag

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, Series B, 34, 187–220.

* Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*. New York: Chapman and Hall.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society*, Series B, 30, 248–275.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

* Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival models and data analysis*. New York: John Wiley & Sons.

* Elber, C. and Ridder, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazards model. *Review of Economic Studies*, 49, 402–411.

Engel, R. F. (1984). Wald, likelihood ratio and Lagrange multiplier tests in econometrics. In Z. Griliches and M. Intrilligator (eds.), *Handbook of Econometrics*. New York: North-Holland.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, 52, 203–223.

* Gross A. J. and Clark, V. A. (1975). *Survival distributions: Reliability applications in the biomedical sciences*. New York: John Wiley & Sons.

* Han, A. and Hausman, J. (1986). *Semiparametric estimation of duration and competing risks models*. Department of Economics, Massachusetts Institute of Technology, Cambridge, Mass.

* Heckman, J. and Singer, B. (1984). The identifiability of the proportional hazards model. *Review of Economic Studies*, 51, 321–341.

* Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52, 271–320.

Hocking, R. R. (1983). Developments in linear regression methodology: 1959-82. *Technometrics*, 25, 219-230.

Hosmer, D.W., Jr. and Lemeshow, S. (2002). *Applied Survival Analysis Regression modeling of time to event data*. Hoboken, N.J.: Wiley-Interscience.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 75-83.

Kalbfleisch, J. and Prentice, R. (2002). *The statistical analysis of failure time data*. 2nd ed. Hoboken, N.J.: Wiley-Interscience.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Lagakos, S. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 35, 139–56.

* Lancaster, T. (1985). Generalized residuals and heterogeneous duration models: With applications to the Weibull model. *Journal of Econometrics*, 28, 155–169.

* Lancaster, T. (1988). *Econometric analysis of transition data*. Cambridge: Cambridge University Press.

Lawless, J. F. (2002). *Statistical models and methods for lifetime data*. Hoboken, N.J.: Wiley-Interscience.

Lee, E. T. (1980). *Statistical methods for survival data analysis*. Belmont, Calif.: Wadsworth.

* Lee, E.T., Wang, J.W. (2003). *Statistical methods for survival data analysis*. Hoboken, N.J.: Wiley-Interscience.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

Manton, K. G., Stallard, E., and Vaupel, J. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*, 81, 635–644.

Miller, R. (1981). *Survival analysis*. New York: John Wiley & Sons.

Nelson, W. (1978). Life data analysis for units inspected once for failure. *IEEE Transactions on Reliability*, R-27, 4, 274–279.

*Nelson, W. (2003) *Applied life data analysis*. New York: John Wiley & Sons.

Parmar, M. K. B. and Machin, D. (1995). *Survival analysis: A practical approach*. New York: John Wiley & Sons.

Peduzzi, P. N., Hofford, T. R., and Hardy, R. J. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, 36, 511–516.

Peto, R. (1973). Experimental survival curves for interval censored data. *Applied Statistics*, 22, 86-91.

Prentice, R. L. and Kalbfleisch, J. D. (1979) Hazard rate models with covariates. *Biometrics*, 35, 25–39.

* Preston, D. and Clarkson, D. B. (1983). SURVREG: A program for the interactive analysis of survival regression models. *The American Statistician*, 37, 174.

Rao, C. R. (1973). *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley & Sons.

*Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, 461-464.

* Steinberg, D. and Monforte, F. (1987). Estimating the effects of job search assistance and training programs on the unemployment durations of displaced workers. In K. Lang and J. Leonard (eds.), *Unemployment and the Structure of Labor Markets*. London: Basil Blackwell.

Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, 64, 156–160.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, Series B, 38, 290–295.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–454.

Wang, M. (1987). *Nonparametric estimation of survival distributions with interval censored data*. John Hopkins University, Baltimore, Md.

* White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11, 95–103.

(*indicates additional references)

Chapter

14

Test Item Analysis

Herb Stenson

TESTAT provides classical analysis and logistic item-response analysis of tests that are composed of responses to each of a set of test items (variables) by each of a set of respondents (cases). Classical analysis provides test summary statistics, reliability coefficients, standard errors of measurement for selected score intervals, item analysis statistics, and summary statistics for individual cases. Graphical as well as numerical displays are provided.

You also can score individual items for each respondent provided that the test items are of the "right versus wrong" variety. However, TESTAT is not limited to these kinds of data; it will accept and analyze any sort of numerical variables that can be used in SYSTAT. Thus, data from true-false tests, multiple-choice tests, rating scales, physiological measures, etc., can all be analyzed with TESTAT using the classical test theory model.

Analysis using logistic, item-response theory is implemented in TESTAT using an iterative, maximum likelihood procedure to estimate item difficulties, item discrimination indices, and subjects' abilities.

Either a one or two-parameter logistic model can be selected. Item histograms can be printed to examine the fit of each item to the model. TESTAT can save total subject scores into a SYSTAT file.

If you use BY, a test can be analyzed for any subgroups of respondents (cases) that you specify. You also have the option of specifying subsets of items (variables) as a subtest to be analyzed. TESTAT also can save item difficulties and discrimination indices into a file for item banking.

Resampling procedures are available in this feature.

Statistical Background

The two statistical approaches to analyzing data from psychological and educational tests have been termed **classical** and **latent trait**. The classical model assumes that items are imperfect measurements of an underlying factor. Like common factor analysis, a single theoretical (unobserved) factor is assumed to comprise a "true" source of variation and random error accounts for the remaining variation in observed scores. Since we cannot observe this true factor, we can estimate it by making assumptions that the random errors are independent and, usually, normally distributed. Thus, the sum of the item scores can yield an estimate of the true score.

The classical model has no role for items of different difficulty. Indeed, it is assumed that any differences in responding to items is due to the ability of the subjects and not to the difficulty of the items. Consequently, tests developed under the classical model tend to have banks of items all of a similar average difficulty or response pattern.

The latent trait model, on the other hand, postulates an underlying distribution that relates item responses to a theoretical trait. This distribution is usually (as in TESTAT) assumed to be logistic, but it can take other forms. In its parameterization, the latent trait model specifically separates subject abilities (individual differences) and item difficulties (scale differences). Tests developed under the latent trait model tend to have a pool of items that vary in difficulty. Some items are failed (or not endorsed) by most subjects and some are passed (or endorsed) by most subjects. Because of this, a latent trait test is especially well suited for measuring larger ranges of abilities or opinions. In addition, the latent trait model allows a more precise description of the performance of an item than simply the item-test correlation. This helps in screening for poor items in a test.

Because of its more elegant parameterization, the latent trait model is generally regarded by test experts to be superior to the classical model for developing surveys and tests of attributes. Indeed, despite the popularity of the classical model (and its associated statistics such as Cronbach's alpha, item-test correlations, and factor loadings) among nonprofessionals and applied researchers, the latent trait model is the one used by the well-known psychological and educational testing organizations. The continuing popularity of the older classical model may be due to its relative simplicity and the lack of availability of latent trait software in the major statistical packages. Until SYSTAT introduced latent trait modeling in a general statistical package, it was confined to specialized software available at selected academic and commercial sites. SYSTAT offers both methods, but we strongly recommend that you learn and apply the latent trait model to develop tests that you intend to reuse.

Classical Model

The principal statistics in the classical model are reliability measures that represent how well a set of items relate to each other (assuming that they all measure a common factor). The reliability, or internal-consistency, coefficients that are produced by TESTAT are the coefficient of correlation between the odd and even test scores, the Spearman-Brown coefficient based on the odd-even correlation, the Guttman-Rulon coefficient, coefficient alpha for all items, coefficient alpha for odd-numbered items, and coefficient alpha for even-numbered items.

The Spearman-Brown coefficient is based on the assumption that the two halves of the test are strictly parallel. The Guttman-Rulon coefficient is based on the assumption that the two halves are parallel in every sense except for having different variances. We call it the Guttman-Rulon coefficient here because the two different formulas for computing it proposed by Guttman (1945) and Rulon (1939) are algebraically equivalent. Coefficient alpha is the internal consistency measure proposed by Cronbach (1951). It is algebraically equivalent to Formula 20 by Kuder and Richardson (KR20) when the test data are dichotomously scored items.

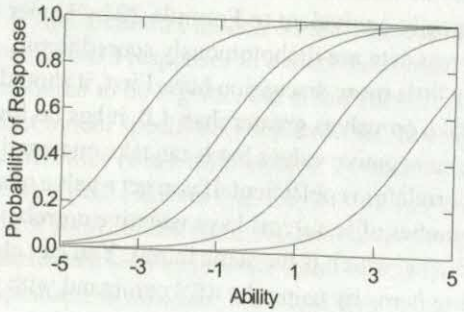
Coefficient alpha deserves a little more discussion here. First, it should be noted that while this coefficient cannot take on values greater than 1.0, it has no lower limit. Therefore, it not only can take on negative values but it can take on negative values less than –1.0, unlike the Pearson correlation coefficient. If you get a value of alpha less than 0, it is because a substantial number of test items have negative correlations with the total test score (or with other items, which is the same thing). You can check the effect of reverse scoring the offending items by using the KEY command with the + and – option (as described later).

Second, a version of alpha called **standardized alpha** is often computed. This coefficient reflects the average size of item-total correlations as opposed to item-total covariances. TESTAT does not produce it for the following reasons. Alpha can be interpreted as the lower limit of reliability for a test that is scored by summing the item scores. If standardized alpha is computed, this coefficient is the lower limit of reliability for a test that is scored by first converting the scores for each item so that all items have equal variances, and then summing these converted scores. Thus, this latter version of alpha does not accurately describe your test unless the items have equal variances. If you need this coefficient, you could first use the STANDARDIZE command to convert each of your items to *z* scores and then run these data with TESTAT. The total scores on the test will then be appropriate for the alpha that is computed, which will be the so-called standardized alpha.

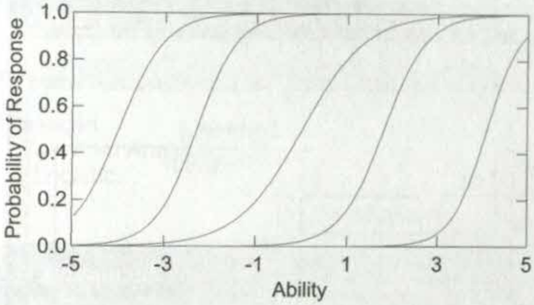
More information about all of these test statistics can be found in standard textbooks such as those by Allen and Yen (1979) and Crocker and Algina (1986).

Latent Trait Model

The latent trait model assigns a probability distribution to responses to each item. Usually, this is a logistic distribution, but it also can be normal. The following figure shows distributions for five hypothetical items. Each curve displays the probability of a correct response on each item by students of different levels of ability. Each item has a common shaped curve based on the cumulative logistic (or normal) distribution function. The only parameter distinguishing the curves is their location. Easier items appear to the left and more difficult items appear to the right. The model generating this graph is called the one-parameter, or Rasch, model.



Often, it is more plausible to assume that items vary in discrimination as well as in difficulty. Items with steeper curves discriminate between subjects of different ability more effectively than items with shallower curves. Not surprisingly, this is called a two-parameter model. The following figure shows an example for five hypothetical items. Notice that the second and third items from the left differ noticeably in discrimination as well as in difficulty.



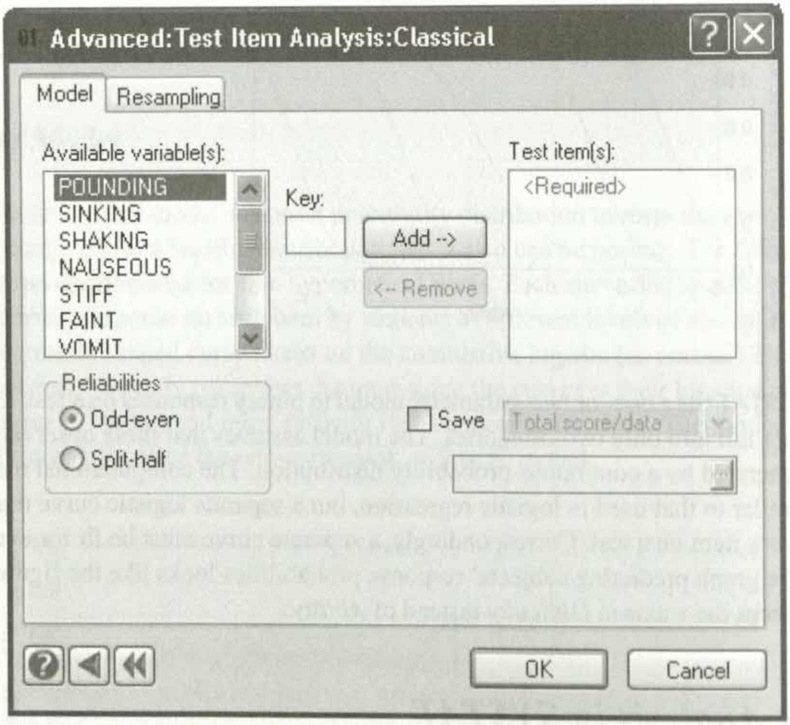
TESTAT fits a one- or two-parameter model to binary responses on a test. The observed data fall into only two categories. The model assumes that these observations were generated by a continuous probability distribution. The computational machinery is similar to that used in logistic regression, but a separate logistic curve must be fit for every item on a test. Correspondingly, a separate curve must be fit for every subject. The graph predicting subjects' response probabilities looks like the figures above, except the *x* axis is *Difficulty* instead of *Ability.*

Test Item Analysis in SYSTAT

Classical Test Item Analysis Dialog Box

To open the Classical Test Item Analysis dialog box, from the menus choose:

Advanced
  Test Item Analysis
    Classical...



Model selection and estimation are available in the Model tab of the Classical Test Item Analysis dialog box.

Test item(s). Select a set of test items and move these into the Test item(s) list.

Key. You can alter the nature of the data by scoring each item response as correct or incorrect or by reversing the scoring scale. For each variable, enter a scoring key value.

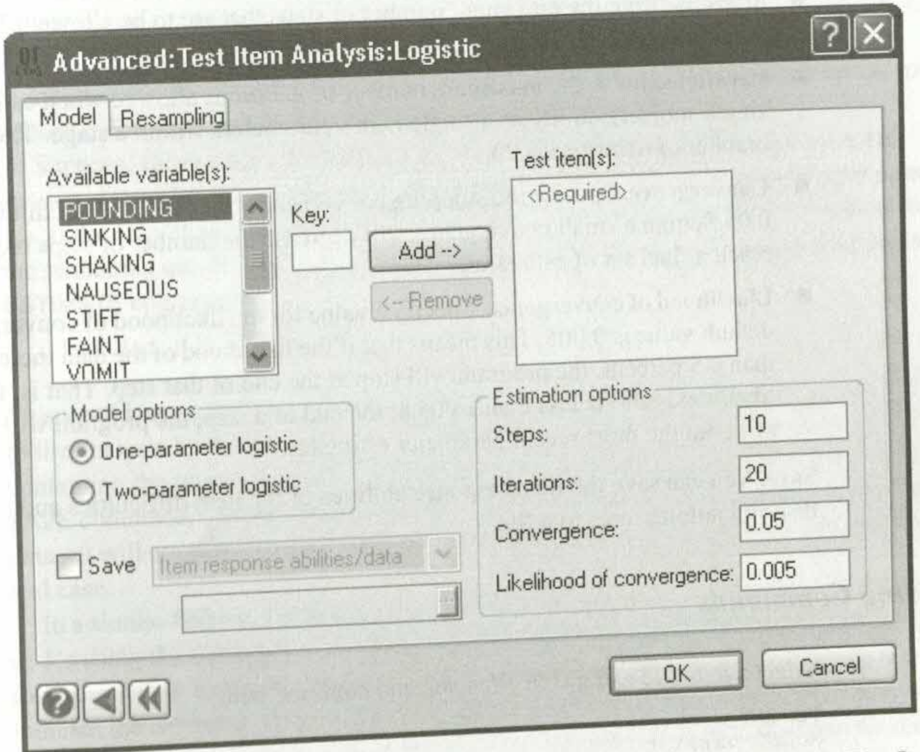
Reliabilities. By default, reliabilities and summary statistics are based on an Odd-even split. Instead of using the Odd-even split, you can select Split-half to use the first half of the items versus the last half of items.

Save. You can save the total subject scores or the average item scores to a file.

Logistic Test Item Analysis Dialog Box

To open the Logistic Test Item Analysis dialog box, from the menus choose:

Advanced
  Test Item Analysis
    Logistic...



Model selection and estimation are available in the Model tab of the Logistic Test Item Analysis dialog box.

If your data values are binary and are coded as zeros and ones, you can analyze your data using item-response theory with the logistic function as the item characteristic curve.

Test item(s). Select a set of test items and move these into the Test item(s) list.

Key. You can alter the nature of the data by scoring each item response as correct or incorrect or by reversing the scoring scale. For each variable, enter a scoring key value.

Model options. Choose between a one-parameter or a two-parameter model. Suppose you select One-parameter logistic, the item discrimination index will be the same for every item, but may change values during the iterative process due to rescaling of the abilities. If you select Two-parameter logistic, each item can have a different discrimination index.

Estimation options. The following can be specified:

- **Steps.** Indicate the maximum number of steps that are to be allowed. The default number of steps is 10.
- **Iterations.** Enter the maximum number of iterations allowed when estimating a single subject's ability or a single item's parameters within a stage. The default number of iterations is 20.
- **Convergence.** Specify the stopping convergence criterion; the default is 0.05.Setting a small convergence will decrease the number of steps required to reach a final set of estimates.
- **Likelihood of convergence.** Specify a value for the likelihood of convergence. The default value is 0.005. This means that if the likelihood of the data increase by less than 0.5 percent, the program will stop at the end of that step. That is, if the likelihood ratio is less than 1.005 at the end of a step, the program will stop and print out the most recent parameter estimates.

Save. You can save the item response abilities or the item difficulties and discrimination scores to a file.

Using Commands

Select a data file using USE *filename* and continue with:

```
TESTAT
MODEL varlist
KEY values
SAVE filename / ITEM
ESTIMATE / CLASSICAL HALF (or)
         / LOG1 or  LOG2, STEPS=n1 ITER=n2 CONVERGE=d1 LCONVERGE=d2
         SAMPLE =BOOT(m,n) SIMPLE(m,n) JACK
```

Usage Considerations

Types of data. By default, TESTAT will use whatever data are in the data set to perform the analyses. However, if you want to alter the nature of these data by scoring each item response as correct or incorrect, or by reversing the scoring scale, you can use KEY. It has two forms.

The first form is used as a scoring key to score each item response as a 0 or a 1, which can mean "incorrect" and "correct," or any other meaningful binary designation. To use this form, you must provide the scoring key as a sequence of non-negative numbers corresponding in a one-to-one fashion to the sequence of items on the test (or subtest). The numbers in your data set must not be negative.

Suppose, for example, that your data set consists of five questions that must be answered "true" or "false" and that you have coded the respondents' answers as 0's and 1's. If the correct answer to questions 1, 2, and 4 is "true" , then the correct answers for the remaining questions is "false", then you would type the KEY command prior to the ESTIMATE command as follows:

```
KEY 1,1,0,1,0
```

This would cause the item responses to be scored according to your scoring key prior to the analysis by the ESTIMATE command. If you want to create a SYSTAT data set containing the scored items, you must also precede the ESTIMATE command with a SAVE command, naming the data set into which the scored data are to be saved. This data set will contain 1's and 0's indicating correct or incorrect responses for each item and case.

In a similar fashion, the responses to multiple-choice items can also be scored as 0's or 1's using the scoring key. If, for example, your five-question test was made up of four-alternative multiple-choice items, then you can use the numbers 1 through 4 to indicate the correct answers in the scoring key. Of course, the respondents' answers must also be entered into the input data set as the numbers 1, 2, 3, or 4. Suppose that your input data set containing responses to the five questions was named *MYFILE*. Then the following commands would score these data as 0's and 1's, save them into a SYSTAT data set named *SCORED*, and produce test and item histograms:

```
USE MYFILE
TESTAT
MODEL var1,var2,var3,var4,var5
SAVE SCORED
KEY 3,1,4,2,1
ESTIMATE
```

The data set that is saved (*SCORED* in this case) will contain as an extra variable the total score for each subject (case).

The second form of the KEY command is used to reverse the scoring of selected items. It can be used when the largest data values for one item indicate the same thing as the smallest data values for another item. This scoring key consists of a sequence of "+" and "–" signs to indicate that the item scores are to be multiplied by a +1 or –1, thus reversing the direction of scoring in the case of –1. Reversing the scoring scale in this way will not affect item variances, but it will alter the possible ranges of item means and total scores. Thus, you might use this method to check the effect on alpha of reversing one or more items. If this increases alpha and it makes sense in the context of the test, you may want to use the LET command to change the scoring of such items so that the highest response score possible would be replaced by the lowest response score possible and so on for the lowest and any intermediate responses.

For example, the following commands will save the input from five items, multiplied by their corresponding weights of +1 and –1, into the data set *WEIGHT*, and produce the default output of ESTIMATE using first-half, last-half as the Split-half option:

```
SAVE WEIGHT
KEY +,-,-,+,+
ESTIMATE / HALF
```

Print options. The default output statistics consist of summary statistics for the test and a set of reliability (internal consistency) coefficients. The output statistics are the mean, standard deviation, standard error of the mean, maximum and minimum values, and the number of cases on which these were computed for the following summary variables: total score (summed across the variables), total score/number of items, total score on odd-numbered items, and total score on even-numbered items.

If the total number of items (variables) in the data set is odd, then the total score for odd-numbered items will be based on one more item than the total score for even-numbered items.

Note that the standard deviations, in keeping with tradition in test theory, are based on sums of squared deviations divided by N, rather than $N-1$. To give unbiased estimates, the standard errors of means are computed by dividing the standard deviations by the square root of $N-1$.

If you want to see item analysis statistics in addition to the test statistics, use PLENGTH LONG.

The first set of additional data that will be provided when you use PLENGTH LONG is the approximate standard error of measurement for total test scores in each of 15

score intervals. These intervals are each 1/2 standard deviation wide and are centered at the mean. Thus, they are the so-called Stanine intervals. The intervals are shown in both z score and total score metrics, so that, even if you have no need for these standard errors of measurement, the table will be useful for seeing how various total scores translate into z scores.

The standard error of estimate shown for an interval is the square root of the average squared difference between odd and even scores (or first minus last half for cases whose total score is in the interval). This is a method recommended by Livingston (1982) and studied empirically by Lord (1984). Lord showed that standard errors of estimate computed by Livingston's method approximate the standard errors of estimate that he got using a three-parameter logistic model to analyze a large set of achievement test data. However, Lord cautions against the use of these estimates if the number of cases in an interval is small or if the interval is near the minimum or maximum total score that is possible.

The second set of additional data that is provided when you use PLENGTH LONG is a set of item statistics that are useful in performing an item analysis of a test. Shown for each item are the item mean and standard deviation, the correlation of the item with the total score, the item reliability index (item-total correlation times standard deviation), the item-total correlation if the item is excluded from the total, and the value of the coefficient alpha if the item is excluded from the test.

Quick Graphs. If your input data are binary "right versus wrong" data, each item plot shows the percentage of the cases in a z-score interval that got the item correct. That is, the axis labeled *Scaled Mean-Item Score* shows the percentage correct for each interval. However, if your data are not of the "right versus wrong" variety, then the *Scaled Mean-Item Score* is the mean-item score for cases in an interval, scaled so that its minimum possible value is 0 and its maximum possible value is 100. (The minimum and maximum values are found by locating the largest and smallest data values that exist in the input data.) Note that the N and percentage listed next to the histograms are the number of cases and percentage of cases with scores in an interval, not the percentage correct. The column labeled *SCORE* gives the actual score.

For the latent trait models, Quick Graphs of the fitted logistic curves are plotted for each item in a grouped array.

Saving files. You can save average item scores ("difficulties" or, in the case of binary items, *p-values*) into a SYSTAT file. The file will include, on each record, the name of an item and its average score.

BY groups. TESTAT analyzes data by groups.

Case frequencies. TESTAT uses the FREQ variable, if present, to duplicate cases. This inflates the total degrees of freedom to be the sum of the number of frequencies. Using a FREQ variable does not require more memory, however.

Case weights. TESTAT weights sums of squares and cross products using the WEIGHT variable for rectangular data input. It does not require extra memory.

Examples

Example 1
Classical Test Analysis

The following data are reports of fear symptoms selected by United States soldiers after being withdrawn from World War II combat. The data were originally reported by Suchman in Stouffer et al. (1950). The variable *COUNT* contains the number of soldiers in each profile of symptom reports.

Notice that we use the FREQ command to implement the case weighting variable *COUNT*. TESTAT weights the cases according to this count before computing statistics. We also save the estimates.

The input is:

```
TESTAT
  USE COMBAT
  MODEL POUNDING..URINE
  FREQ COUNT
  IDVAR COUNT
  SAVE TEMP / ITEM
  ESTIMATE / CLASSICAL
```

The output is:

```
SYSTAT Rectangular file contains variables:

POUNDING     SINKING    SHAKING    NAUSEOUS    STIFF    FAINT
VOMIT        BOWELS     URINE      COUNT

Case Frequencies Determined by Value of Variable COUNT

Data Below are Based on 93 Complete Cases for 9 Data Items

 Test Score Statistics

                     | Total    Average     Odd     Even
---------------------+----------------------------------
Mean                 | 4.538      0.504   2.473    2.065
```

Standard Deviation	2.399	0.267	1.333	1.277
Standard Error	0.250	0.028	0.139	0.133
Maximum	9.000	1.000	5.000	4.000
Minimum	1.000	0.111	0.000	0.000
N of Cases	93	93	93	93

Internal Consistency Data

Split-half Correlation	: 0.690
Spearman-Brown Coefficient	: 0.816
Guttman (Rulon) Coefficient	: 0.816
Coefficient Alpha - All Items	: 0.787
Coefficient Alpha - Odd Items	: 0.613
Coefficient Alpha - Even Items	: 0.661

Approximate Standard Error of Measurement of Total Score for 15 z score Intervals

z-score	Total Score	N	Standard Error
-3.750	-4.458	0	.
-3.250	-3.258	0	.
-2.750	-2.059	0	.
-2.250	-0.860	0	.
-1.750	0.340	10	1.000
-1.250	1.539	16	1.000
-0.750	2.739	6	1.000
-0.250	3.938	29	1.390
0.250	5.137	10	1.095
0.750	6.337	8	1.000
1.250	7.536	8	0.000
1.750	8.735	6	1.000
2.250	9.935	0	.
2.750	11.134	0	.
3.250	12.334	0	.

Item Reliability Statistics

Item	Label	Mean	Standard Deviation	Item Total R	Item Reliability Index	Excl Item R
1	POUNDING	0.903	0.296	0.331	0.098	0.215
2	SINKING	0.785	0.411	0.499	0.205	0.354
3	SHAKING	0.559	0.496	0.678	0.336	0.539
4	NAUSEOUS	0.613	0.487	0.721	0.351	0.599
5	STIFF	0.538	0.499	0.693	0.346	0.559
6	FAINT	0.452	0.498	0.715	0.356	0.588
7	VOMIT	0.376	0.484	0.622	0.301	0.472
8	BOWELS	0.215	0.411	0.625	0.257	0.502
9	URINE	0.097	0.296	0.503	0.149	0.402

Item Reliability Statistics (contd...)

Excl Item Alpha
0.794
0.782
0.757
0.747
0.754
0.749
0.767
0.763
0.777

Use PLENGTH LONG to see item histograms for this test.

Example 2
Logistic Model (One Parameter)

If your data are binary and are coded as 0's and 1's or recoded with the KEY command, you can analyze your data using item-response theory with the LOGISTIC function as the item characteristic curve. Either a one-parameter (Rasch) model or a two-parameter logistic model can be implemented by using the MODEL command. The one-parameter model is the default.

The input is:

```
TESTAT
  USE COMBAT
  MODEL POUNDING..URINE
  FREQ COUNT
  IDVAR COUNT
  SAVE TEMP / ITEM
  ESTIMATE / LOG1
```

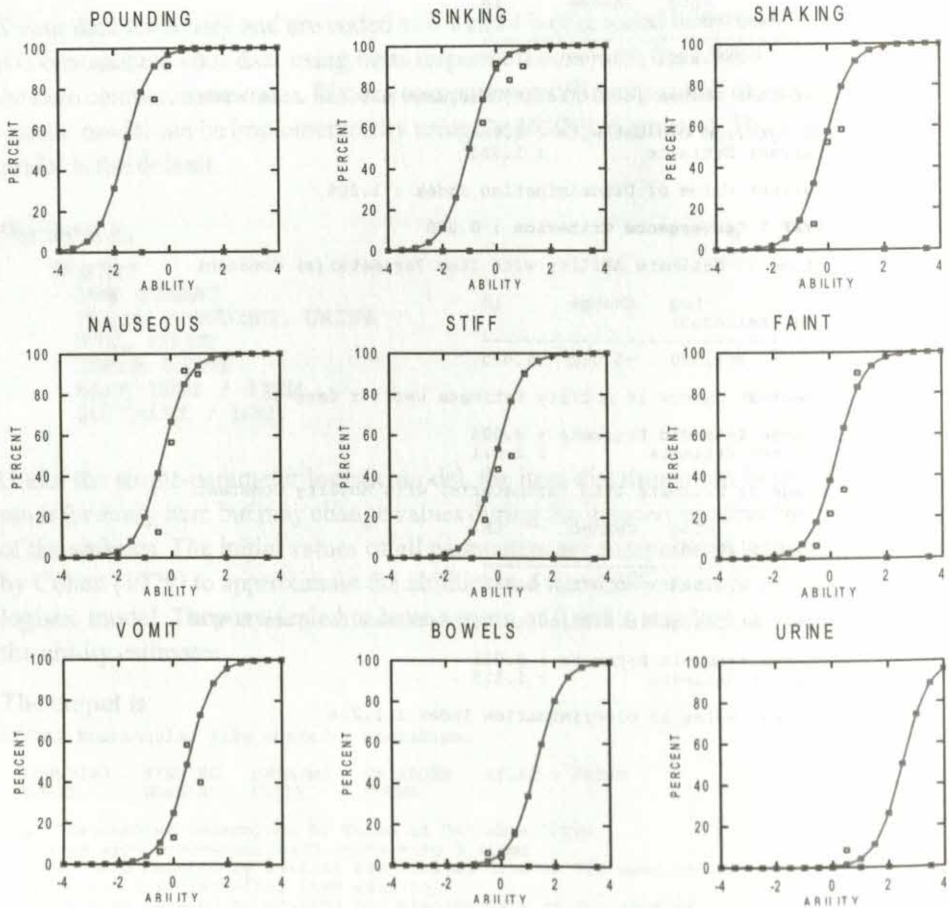
Under the single-parameter logistic model, the item discrimination index will be the same for every item but may change values during the iterative process due to rescaling of the abilities. The initial values of all parameters are computed by a technique given by Cohen (1979) to approximate the abilities and item difficulties of a one-parameter logistic model. They are scaled to have a mean of 0 and a standard deviation of 1 for the ability estimates.

The output is:

```
SYSTAT Rectangular file contains variables:

  POUNDING   SINKING   SHAKING   NAUSEOUS   STIFF   FAINT
  VOMIT      BOWELS    URINE     COUNT

Case Frequencies Determined by Value of Variable COUNT
93 Cases were processed, each containing 9 items
6 Cases were deleted by editing for missing data or for zero or
perfect total scores after item editing.
0 Items were deleted by editing for missing data or for zero or
perfect total scores after item editing.

Data below are based on 87 Cases and 9 Items

Total Score Mean                                      :   4.230
Standard Deviation                                    :   2.164
-Log(Likelihood) Using Initial Parameter Estimates : 270.982

STEP 1 Convergence Criterion : 0.050

Stage 1: Estimate Ability with Item Parameter(s) Constant

          -Log    Change      LR
  (Likelihood)
------------------------------
       270.071   -0.911   2.486
Greatest Change in Ability Estimate was for Case 80
```

```
Change from Old Estimate : 0.134
Current Estimate         : 2.005

Stage 2: Estimate Item Parameter(s) with Ability Constant

      -Log   Change      LR
 (Likelihood)
 -------------------------------
     269.662  -0.409   1.505

Greatest Change in Difficulty Estimate was for Item BOWELS

Change from Old Estimate : 0.084
Current Estimate         : 1.301

Current Value of Discrimination Index : 1.206

STEP 2 Convergence Criterion : 0.050

Stage 1: Estimate Ability with Item Parameter(s) Constant

      -Log   Change      LR
 (Likelihood)
 -------------------------------
     269.590  -0.072   1.075

Greatest Change in Ability Estimate was for Case 87

Change from Old Estimate : 0.006
Current Estimate         : 2.011

Stage 2: Estimate Item Parameter(s) with Ability Constant

      -Log   Change      LR
 (Likelihood)
 -------------------------------
     269.549  -0.041   1.042

Greatest Change in Difficulty Estimate was for Item BOWELS

Change from Old Estimate : 0.032
Current Estimate         : 1.315

Current Value of Discrimination Index : 1.226
```

Latent Trait Model Item Plots



Three levels of the iterative process must be distinguished here. The program operates in what are labeled *STEPS* in the output. Each step consists of two stages. In stage 1, the subjects' abilities are estimated, one at a time, holding the item parameters constant at their most recent values. At the end of stage 1, the resulting abilities are rescaled so as to have a mean of 0 and a standard deviation of 1. The item parameters are also rescaled to conform to the new ability scale. In stage 2, the item parameter(s) are estimated, one item at a time, holding the abilities constant at their most recent values. A new step is then begun, if necessary, in which this two-stage process is repeated.

Within each stage is the third level of the iterative process, called *ITER*. Here, as a single ability (in stage 1) or a single item's parameters (in stage 2) are being estimated, successive iterations are performed until the parameter being estimated does not change by more than a tolerance value called *Tol*. When this criterion is met, the program moves on to estimate the ability for the next case (in stage 1) or the next item's parameters (in stage 2). This iterative process is repeated within a stage until the data are exhausted. Then the next stage is begun.

There are two criteria for stopping the step-wise process. At the end of each stage, the likelihood of obtaining the test data that are in the input data set is computed, given the current values of all parameters. The negative logarithm of this likelihood and the change in this value from the previous stage are printed. The ratio of the current likelihood to the previous is also computed and printed. If, at the end of a step (after stage 2), this likelihood ratio is less than a value specified by a stopping criterion called LCONVERGE, no further steps are run, and the final item parameters are printed. This is the first stopping criterion.

The second stopping criterion relies on the maximum change in parameter estimates between stages. At the end of a stage, the maximum change in the parameters being estimated in that stage is printed. If, at the end of a step, no parameter estimated in either stage of that step changed more than the value of CONVERGE, the stage-wise process is terminated, and the final item parameters are printed. Thus, the program will stop entering new steps whenever either of the two stopping criteria is met, whichever occurs first.

The final parameter estimates are, thus, a type of maximum likelihood estimate. However, you should realize that because the process alternates between estimating item parameters and abilities, the final parameter estimates are not a true maximum likelihood estimate of all parameters simultaneously. As with other programs that use this same type of alternating estimation technique, the process does converge for all but very unusual data sets.

Example 3
Logistic Model (Two Parameter)

The 20-item version of the Social Desirability Scale described by Strahan and Gerbasi (1972) was administered as embedded items in another test to 359 undergraduate students in psychology. The social desirability items were scored for the "social desirability" of the response and coded as 0's and 1's in a SYSTAT data set named *SOCDES*.

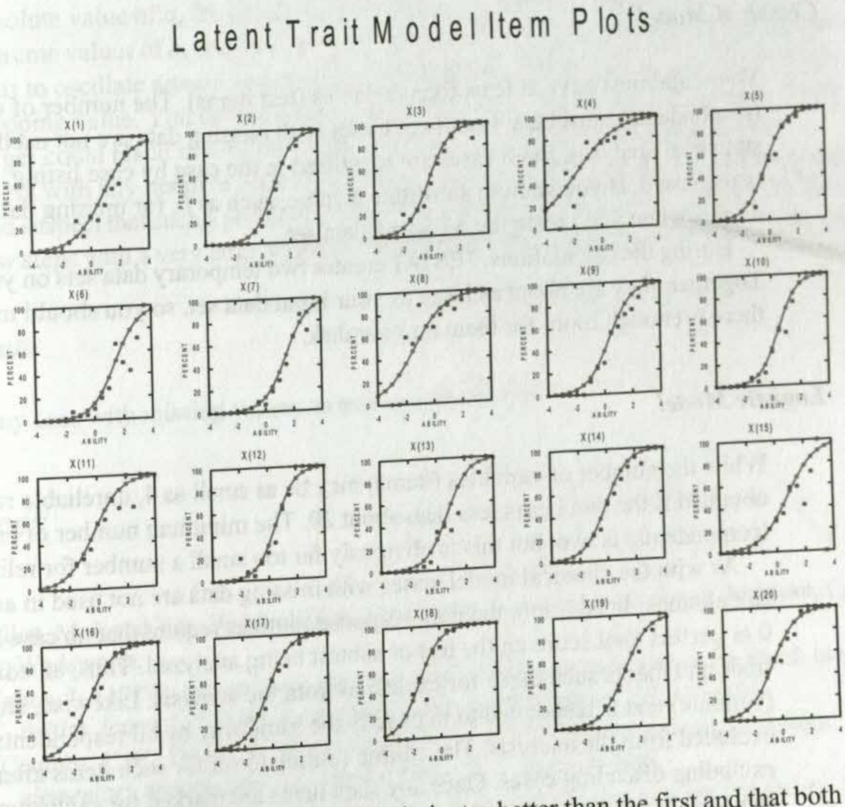
The input is:

```
TESTAT
  USE SOCDES
  MODEL X(1..20)
  SAVE TEMP / ITEM
  ESTIMATE / LOG2, STEP=2, CONVERGE=0.1
```

The output is:

```
SYSTAT Rectangular file contains variables:
  X(1)     X(2)     X(3)     X(4)     X(5)     X(6)
  X(7)     X(8)     X(9)     X(10)    X(11)    X(12)
  X(13)    X(14)    X(15)    X(16)    X(17)    X(18)
  X(19)    X(20)

359 Cases were processed, each containing 20 items
4 Cases were deleted by editing for missing data or for zero or
perfect total scores after item editing.
0 Items were deleted by editing for missing data or for zero or
perfect total scores after item editing.
Data below are based on 355 Cases and 20 Items

Total Score Mean                                     :    9.386
Standard Deviation                                   :    3.992
-Log(Likelihood) Using Initial Parameter Estimates : 3634.928

STEP 1 Convergence Criterion : 0.100

Stage 1: Estimate Ability with Item Parameter(s) Constant

       -Log     Change      LR
  (Likelihood)
 ------------------------------
     3634.122   -0.806   2.239

Greatest Change in Ability Estimate was for Case 139

Change from Old Estimate : -0.106
Current Estimate         :  2.956

Stage 2: Estimate Item Parameter(s) with Ability Constant

       -Log      Change          LR
  (Likelihood)
 -------------------------------------
     3622.570   -11.552   104021.460

Greatest Change in Difficulty Estimate was for Item X(19)

Change from Old Estimate : 0.022
Current Estimate         : 0.947

Greatest Change in Discrimination Estimate was for Item X(8)

Change from Old Estimate : -0.164
Current Estimate         :  0.531

STEP 2 Convergence Criterion : 0.100

Stage 1: Estimate Ability with Item Parameter(s) Constant

       -Log     Change      LR
  (Likelihood)
 -------------------------------
     3619.923   -2.647   14.116
```

```
Greatest Change in Ability Estimate was for Case 66

Change from Old Estimate : -0.181
Current Estimate         : -2.266

Stage 2: Estimate Item Parameter(s) with Ability Constant

       -Log    Change        LR
  (Likelihood)
 ----------------------------------
     3612.343   -7.579   1957.627

Greatest Change in Difficulty Estimate was for Item X(4)

Change from Old Estimate : -0.581
Current Estimate         : -1.770

Greatest Change in Discrimination Estimate was for Item X(4)

Change from Old Estimate : -0.153
Current Estimate         :  0.408
```



You can see that the second item discriminates better than the first and that both items seem to fit the model moderately well.

Computation

All calculations are performed, with provisional algorithms used for calculating all means and sums of squares that are needed. The formulas for all of the statistics that are shown in the output can be found in the references given below.

Algorithms

Provisional algorithms are used for means and sums of squares. The calculations for the classical and logistic models are as follows:

Classical Model

Your data must have at least four variables (test items). The number of cases (respondents) must be at least two. Cases with missing data are not used in any of the statistical analyses. Such cases are identified in the case by case listing, if this listing is requested. If you want to substitute a value, such as 0, for missing data, you should do this when you create the SYSTAT data set.

During the calculations, TESTAT creates two temporary data sets on your data disk. Together, they are about as large as your input data set, so you should make sure that there is enough room for them on your disk.

Logistic Model

While the number of variables (items) may be as small as 4, unreliable results will be obtained if the number is less than about 20. The minimum number of cases (respondents) is two, but this is obviously far too small a number for reliable results.

As with the classical model, cases with missing data are not used in any of the calculations. In addition, the item-response routines require that no case have either a 0 or perfect total score on the test or subtest being analyzed. Thus, an editing routine finds and marks such cases for exclusion from the analysis. Likewise, any item (variable) that is responded to in exactly the same way by all respondents must be excluded from the analysis. The editing routine looks for such items after first excluding offending cases. Once any such items are marked for exclusion, the routine again looks for inappropriate cases, using only the remaining items. It iterates in this fashion until no inappropriate cases or items remain. Any items or cases that have been

excluded from the analysis are reported by the output routines. The same temporary data sets that are mentioned above for the classical model are also created for the logistic model. Make sure that your disk has room for them.

The algorithm for finding the maximum likelihood (actually, the minimum of the negative logarithm of the likelihood) for each ability in stage 1 and for each item's parameter(s) in stage 2 is based on Fletcher-Powell minimization (Press et al., 1992).

The logistic model that is used in this program is the now-familiar two-parameter formula found in Lord (1980), Hulin et al. (1983), and many other references.

The discrimination parameter for an item is a and the difficulty is b, while the subject's ability is 0. In TESTAT, the function to be minimized is designed to place limits on the values of 0 and a by "driving the iterative routine away" from estimates greater than these limits. The limits are 6.00 for the absolute value of 0 and 3.00 for the absolute value of a, the discrimination index. If your data imply a lot of items with extreme values of a, or a lot of extreme values of 0, you may find that the program will start to oscillate around some value of the likelihood ratio that is not less than the stopping value. You cannot change these limits because to make them very much larger could result in illegally large values for the exponent (x) in the model.

As with any iterative estimation procedure, you should beware of local minima. If you suspect that such a problem exists after inspecting your output, try running the first few steps with a very large value of CONVERGE and then switching to a smaller value.

Missing Data

Any case with missing values on any item is deleted.

References

Allen, M. J. and Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, Calif.: Wadsworth.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 32, 113–120.

*Coombs, C. H., Dawes, R. M., and Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, N.J.: Prentice-Hall.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt Rinehart Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.

Hulin, C. L., Drasgow, F., and Parsons,C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, Ill.: Dow Jones-Irwin.

Livingston, S. (1982). Estimation of conditional standard error of measurement for stratified tests. *Journal of Educational Measurement*, 19, 135–138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Erlbaum.

Lord, F. M. (1984). Standard error of measurement at different ability levels. Technical Report Number RR-84-8. Princeton, N.J.: Educational Testing Service.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical recipes: The art of scientific computing*. 2nd ed. Cambridge: Cambridge University Press.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99–103.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Staf, S. A., and Clausen, J. A. (1950). *Measurement and prediction*. Princeton, N.J.: Princeton University Press.

Strahan, R. and Gerbasi, K. C. (1972). Short, homogeneous versions of the Crowne-Marlowe social desirability scale. *Journal of Clinical Psychology*, 28, 191–193.

(* indicates additional reference)

Chapter

15

Time Series

Leland Wilkinson and Yuri Balasanov
(revised by S M Adinarayana Murthy and Rupesh Kumar)

Time Series implements a wide variety of time series models, including linear and nonlinear filtering, Fourier analysis, seasonal decomposition, nonseasonal and seasonal exponential smoothing, and the Box-Jenkins approach (Box et al., 1994) to nonseasonal and seasonal ARIMA. SYSTAT also provides nonparametric techniques for detecting and estimating trends in a series. You can save results from transformations, smoothing, the deseasonalized series, and forecasts for use in other SYSTAT procedures.

The general strategy for time series analysis is to:

- Plot the series using Time Series Plot, ACF, PACF, or CCF
- Transform the data to stabilize the variance across time or to make the series stationary using Transform
- Smooth the series using moving averages, running medians, or general linear filters using LOWESS or Exponential smoothing
- Fit your model using ARIMA
- Examine the results by plotting the smoothed or forecasted results

Before performing a particular time series analysis, you can specify how missing values should be handled.

- **Interpolate.** Interpolates missing values by using DWLS (Distance Weighted Least-Squares). DWLS interpolates by locally quadratic approximating curves that are weighted by the distance to each nonmissing point in the series. With this algorithm, all nonmissing values in the series contribute to the missing data estimates, and thus complex local features can be modeled by the interpolant.

- **Delete.** Prevents interpolation and only the leading nonmissing values are retained for analysis. In series that begin with one or more missing values, the series is deleted from the first missing value following one or more nonmissing values. This option enables you to forecast missing values from a nonmissing subsection of the series. You can then insert these forecasts into the series and repeat the procedure later in the series if necessary.

Statistical Background

Time series analysis can range from the purely exploratory to the confirmatory testing of formal models. Series encompasses both exploratory and confirmatory methods. Among the exploratory methods are smoothing and plotting. Among confirmatory models are two general approaches: time domain and frequency domain. In time-domain models, we examine the behavior of variables over time directly. In frequency-domain models, we examine frequency (periodic) components contributing to a time series.

Time-domain (autoregressive, moving average, and trend) models represent a series as a function of previous points in the same series or as a systematic trend over time. Time-domain models can fit complex patterns of time series with just a few parameters. Makridakis et al. (1997), McCleary and Hay (1980), and Nelson (1973) introduce these models, while Box et al. (1994) provide the primary reference for ARIMA models.

Frequency-domain (spectral) models decompose a series into a sum of sinusoidal (waveform) elements. These models are particularly useful when a series arises from a relatively small set of cyclical functions. Bloomfield (2000) introduces these models.

In this introduction, we will discuss exploratory methods (smoothing), time-domain models (ARIMA, seasonal decomposition, exponential smoothing), and frequency-domain (Fourier) models.

Smoothing

Smoothing is a complex topic whose applications exceed space here; consult Velleman and Hoaglin (1981) or Bloomfield (2000) for more complete discussions.

Moving Averages

One of the simplest smoothers is a moving average. If a data point consists of a smooth component plus random error, then if we average several points surrounding a point, the errors should tend to cancel each other out.

Here are two possible moving averages three and four points wide. The window shows which points are being averaged. The boldface shows which point in the series is replaced with the average.

Three-point window

Series	y1	y2	y3	y4	y5	y6	y7	y8	y9
Window	y1	y2	y3						
		y2	y3	y4					
			y3	y4	y5				
				y4	y5	y6			
					y5	y6	y7		
						y6	y7	y8	
							y7	y8	y9
New series	y1	x2	x3	x4	x5	x6	x7	x8	y9

Four-point window

Series	y1	y2	y3	y4	y5	y6	y7	y8	y9	*y10*
Window	y1	y2	y3	y4						
		y2	y3	y4	y5					
			y3	y4	y5	y6				
				y4	y5	y6	y7			
					y5	y6	y7	y8		
						y6	y7	y8	y9	
							y7	y8	y9	*y10*
New series	y1	y2	x3	x4	x5	x6	x7	x8	x9	*y10*

Notice that the four-point window does not have a point in the series at its center. Consequently, we replace the right point of the two in the middle with the average of

the four points. This rule is followed for all even windows except two-point windows. Two-point windows can thus be used to shift asymmetrical smoothings back to the left.

If you prefer algebra, then the following description shows how the three-point window smooths *y* into *x*.

$$x_1 = y_1$$
$$x_2 = (y_1 + y_2 + y_3)/3$$
$$x_3 = (y_2 + y_3 + y_4)/3$$

Notice also that the first and last points in the series are unchanged by the three-point window of moving averages. The four-point window leaves the first two and last one points unchanged.

Weighted Running Smoothing

If you know something about filter design (see Bloomfield, 2000), you can construct a more general linear filter by using weights. In the examples, we illustrate seven- and four-point moving averages with equal weights.

The smoothings in the examples used even weights of 1 for each member in the window since we did not specify otherwise. We could, however, set these weights to any real number; for example, 1,2,1. Some of you may recognize these as Hanning weights (Chambers et al., 1983; Velleman and Hoaglin, 1981). It is possible to show algebraically that weighting by (1,2,1) in a three-observation window is the same as smoothing twice with equal weights in a two-observation window. The DWLS smoothing method for graphics is a form of weighting in which weights are determined by distance weighted least-squares.

Running Median Smoothers

Now, let us look at another smoother — running medians. Sometimes it is handy to have a more robust filter when you suspect the data do not contain Gaussian noise. You can choose this filter with the Median option. It works like the Mean option, except the values in the series are replaced by the median of the window instead of the mean.

Can you see why running mean and running median smoothers with a window of two are the same?

We can use combinations of these smoothers to construct more complex nonlinear filters. The following sequence of smoothings comprises a nonlinear filter because it

does not involve a simple weighted average of the values in a window (except for the final Hanning step). It uses a combination of running medians instead:

Running median smoother, window 4
Running median smoother, window 2
Running median smoother, window 5
Running median smoother, window 3
Running means smoother, window 3, weights 1, 2, 1

You can read about this filter (called **4253H**) in Velleman and Hoaglin (1981). It is due to the work of Tukey (1977). It happens to be a generally effective compound smoother because it clears outliers out of the sequence in the early stages and polishes up the smooth later. Velleman and Hoaglin (1981) use this smoother twice on the same data by smoothing the data, smoothing the residuals from this smooth, and adding the two together. You can do this by using Save with the last smoothing to save the smoothed values into a SYSTAT file. You can then merge the files and compute residuals. In the final step, you can smooth the residuals.

LOWESS Smoothing

Cleveland (1979) presented a method for smoothing values of *Y* paired with a set of ordered *X* values. Chambers et al. (1983) introduce this technique and present some clear examples. If you are not a statistician, by the way, and want some background information on recent advances in statistics, read Chambers et al. (1983) (and Velleman and Hoaglin (1981) if you do not know about Tukey's work).

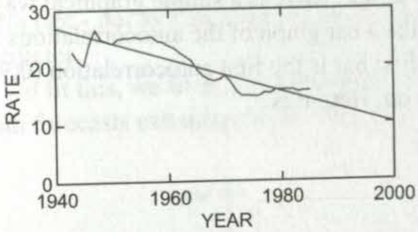
Scatterplot smoothing allows you to look for a functional relation between *Y* and *X* without prejudging its shape (or its monotonicity). The method for finding smoothed values involves a locally weighted robust regression. SYSTAT implements Cleveland's LOWESS algorithm on equally spaced data values. You can also use LOWESS on scatterplots of unequally spaced data values.

ARIMA Modeling and Forecasting

The following data show the U.S. birth rate (per 1000) for several decades during and following World War II. They were compiled from federal statistics, principally the U.S. Census.

YEAR	RATE	YEAR	RATE
1943	22.7	1965	19.4
1944	21.2	1966	18.4
1945	20.4	1967	17.8
1946	24.1	1968	17.5
1947	26.6	1969	17.8
1948	24.9	1970	18.4
1949	24.5	1971	17.2
1950	24.1	1972	15.6
1951	24.9	1973	14.9
1952	25.1	1974	14.9
1953	25.1	1975	14.8
1954	25.3	1976	14.8
1955	25.0	1977	15.4
1956	25.2	1978	15.3
1957	25.3	1979	15.9
1958	24.5	1980	15.9
1959	24.3	1981	15.9
1960	23.7	1982	15.9
1961	23.3	1983	15.5
1962	22.4	1984	15.7
1963	21.7	1985	15.7
1964	21.0		

These data are a **time series** because they comprise values on a variable distributed across time. How can you use these data to forecast birth rates up to the year 2000? A popular statistical method for such a forecast is linear regression. Let us try it. Here is a plot of birth rates against year with the least-squares line. The data points are connected so that you can see the series more clearly.



What's wrong with this forecasting method? You may want to read Chapter 2 of Statistics II (if you haven't already). There, we discussed assumptions needed for estimating a model using least-squares. We can legitimately fit a line to these data by least-squares for the explicit purpose of getting predicted values on the line as close as possible, on the average, to observed values in the data. In forecasting, however, we want to use a fitted model to extrapolate beyond the series. The fitted linear model is:

$$RATE = 579.342 - 0.285 * YEAR$$

If we want our estimates of the slope and intercept in this model to be unbiased, we need to assume that the errors (ε) in the population model are independent of each other and of *YEAR*. Does our data plot give us any indication of this?

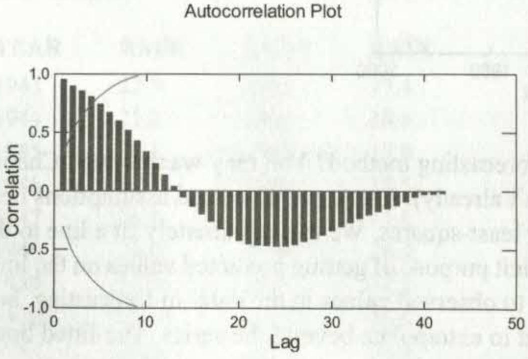
On the contrary, it appears from the data that the randomness in this model is related to *YEAR*. Take any two adjacent years' data. On average, if there is an underprediction one year, there will be an underprediction the next. If there is overprediction one year, there is likely to be overprediction the next. These data clearly violate the assumption of independence in the errors.

Autocorrelation

There is a statistical index that reveals how correlated the residuals are. It is called the **autocorrelation**. The first-order autocorrelation is the ordinary Pearson correlation of a series of numbers with the same series shifted by one observation ($y_2, y_1; y_3, y_2; ...; y_n, y_{n-1}$). In our residuals from the linear model, this statistic is 0.953. If you remember about squaring correlation coefficients to reveal proportion of variance, this means that over 89 percent of the variation in error from predicting one year's birth rate can be accounted for by the error in predicting the previous year's.

The second-order autocorrelation is produced by correlating the series ($y_3, y_1; y_4, y_2; ...; y_n, y_{n-2}$). Computing this statistic involves shifting the series down *two*

years. As you may now infer, we can keep shifting and computing autocorrelations for as many years as there are in the series. There is a simple graphical way to display all these autocorrelations. It looks like a bar graph of the autocorrelations sequenced by year, or index in the series. The first bar is the first autocorrelation (0.953). The next highest bar is the second, and so on. Here it is:



This autocorrelation plot tells us about all the autocorrelations in the residuals from the linear model. As you can see, there is a strong dependence in the residuals. As we shift the series far enough back, the autocorrelations become negative, because the series crosses the prediction line and the residuals become negative. Over the entire series, there are three crossings and three corresponding shifts in sign among the autocorrelations.

Autoregressive Models

We would have the same serial correlation problem if we refined our model to include a quadratic term:

$$\text{RATE} = \beta_0 + \beta_1\text{YEAR} + \beta_2\text{YEAR}^2 + \varepsilon$$

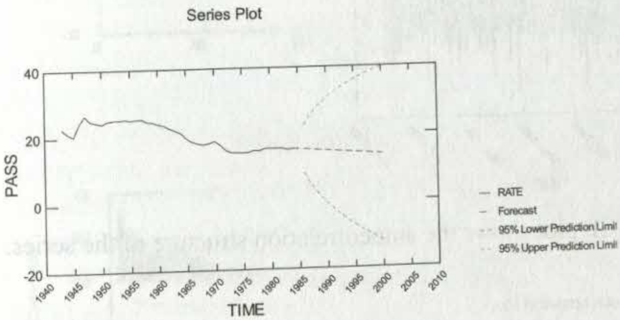
You can try this model with GLM, but you will find a large autocorrelation in the residuals even though the curve fits the data more closely. How can we construct a model that includes the autocorrelation structure itself?

The autoregressive model does this:

$$\text{RATE}_i = \beta_0 + \beta_1\text{RATE}_{i-1} + \varepsilon_i$$

Notice that this model expresses a year's birth rate as a function of the previous year's birth rate - not as a function of *YEAR*. Time becomes a sequencing variable, not a predictor.

To fit this, we fit an AR(1) model with the ARIMA procedure. Here is the result, with forecasts extending to the year 2000:



The forecasted values are represented by the dotted line. Unlike the regression model forecast, the autoregressive forecast begins at the last birth rate value and drifts back toward the mean of the series. This forecast behavior is typical of this particular model, which is often called a **random walk**.

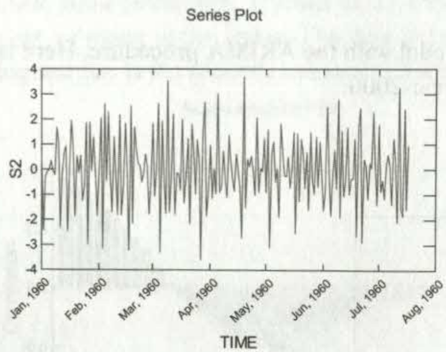
Moving Average Models

There is another series model that can account for fluctuations across time. This models a series as a cumulation of random shocks or disturbances. Consider a time series y_t of daily change in price of a stock. If there is no assignable cause for change in price then we would model y_t simply as $y_t = e_t$, where e_t is a random variable with mean 0. Suppose this stock price is impacted by daily production at a certain oil well and that it takes the market two days to comprehend the impact and react. Then y_{t+2} could consist of its own random component plus a multiple of the change on day (t+1) and could be modeled as:

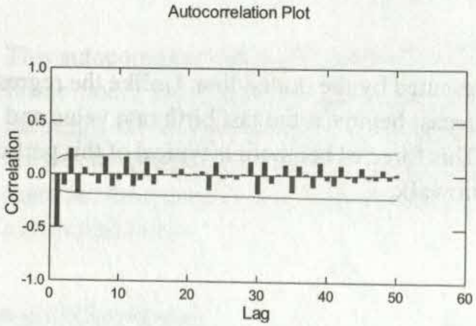
$$y_{t+2} = e_{t+2} + ae_{t+1}$$

Such a series will no longer be a purely random series and will be a Moving Average (MA) series. There will be autocorrelation in the series. An example of such a series for a = -1 is plotted here. This is however a stationary series.

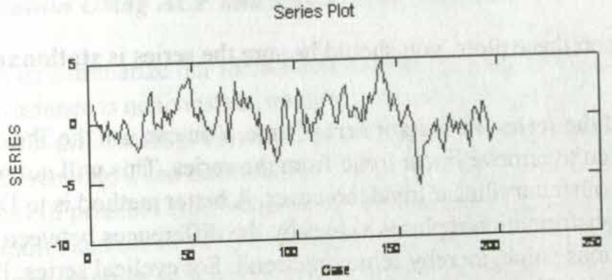
A plot of the daily change in price of a stock is given below:

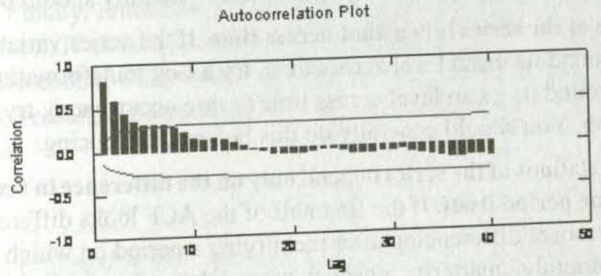


The following ACF plot shows the autocorrelation structure of the series.



Unlike the autoregressive model, which represents an observation as a function of the previous observation's values, the moving average model represents an observation as a function of the previous observation's errors. Here is an example of first-order autoregressive, or AR(1) series, with its autocorrelation function, from which you can observe that its structure is different from that of MA(1).





ARMA Models

Autoregressive and moving average models can be mixed to make autoregressive-moving average models. They can be mixed with different orders, for example, AR(2) plus MA(1), which is often expressed as ARMA(2,1). A text on forecasting will offer instances of these more complicated models. You could visually add the two sample series above, however, to see how an ARMA(1,1) model would look.

Identifying Models

Before you can fit an AR, MA, or ARMA model, you need to identify which model is appropriate for your series. You can look at the series plot to find distinctive patterns, as in the figure contrasting AR(1) and MA(1) directly above. Real data seldom fit these ideal types as clearly, however. There are several powerful tools that distinguish these families of models. We have already seen one: the autocorrelation function plot (ACF). The **partial autocorrelation function plot** (PACF) provides additional information about serial correlation. To identify models, we use both of these plots.

Stationarity

Before working on these plots, you should be sure the series is **stationary**. This implies:

- **The mean of the series is constant across time.** You can use the Trend transformation to remove linear trend from the series. This will not reduce quadratic or other curvilinear trend, however. A better method is to Difference the data. This transformation replaces values by the differences between each value and the previous value, thereby removing trend. For cyclical series, like monthly sales, seasonal differencing may be required before fitting a model (see below). Data that are drifting up or down across the series generally should be differenced.
- **The variance of the series is constant across time.** If the series variation is increasing around its mean level across time, try a Log transformation. If it is decreasing around its mean level across time (a rare occurrence), try a Square transformation. You should generally do this before differencing.
- **The autocorrelations of the series depend only on the difference in time points and not on the time period itself.** If the first half of the ACF looks different from the second, try seasonal differencing after identifying a period on which the data are fluctuating. Monthly, quarterly, seasonal, annual data often cycle this way.

ACF Plots

The autocorrelation function plot displays the pattern of autocorrelations. We have seen in this introduction an ACF plot of the residuals from a linear fit to birth rate. The slow decay of the autocorrelations after the first indicates autoregressive behavior in the residuals.

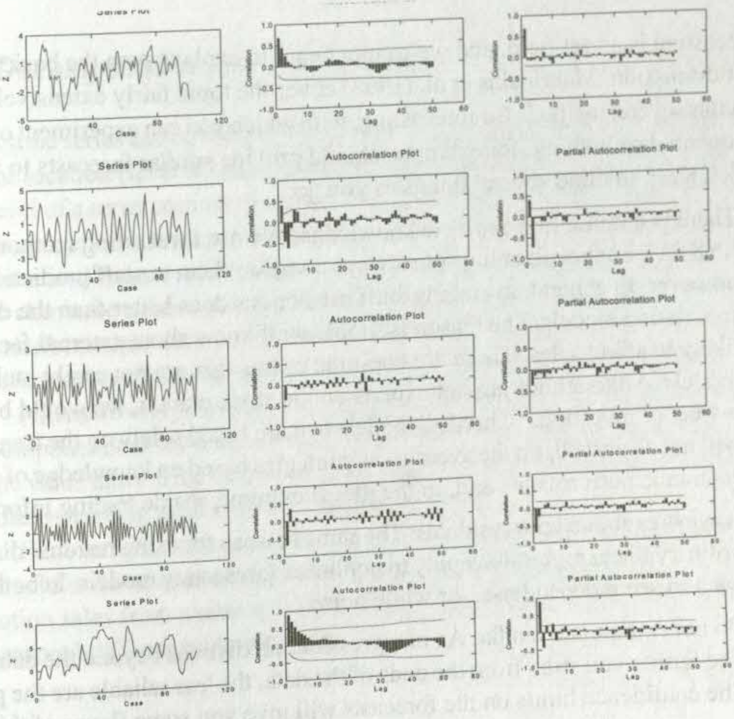
PACF Plots

The partial autocorrelation function plot displays autocorrelations, but each one below the first is conditioned on the previous autocorrelation. The PACF shows the relationship of points in a series to preceding points after partialing out the influence of intervening points. We examine them for effects that do not depend linearly on previous (smaller lag) autocorrelations.

Identification Using ACF and PACF Plots Together

Let us summarize our identification strategy. First, make sure the series is stationary. If variance is nonconstant, transform it with Log or Square. If trend is present, remove it with differencing. Finally, if seasonality is present, remove it with seasonal differencing. Then examine ACF and PACF plots together. On the facing page is a chart of possible types of patterns. To the right of each series is the ACF plot (in the middle) and the PACF plot (on the right). However, because coefficients in the model can be negative or positive, other representations are possible. The plots shown here are based on one particular combination of signs.

Finally, remember that differencing can remove both trend and autoregressive effects. If an AR(1) model fits your data, as with our birth rate example, then differencing will produce only white noise and your ACF will look uniformly random. As a result, differencing is like constraining an autoregressive parameter to be exactly one.



AR(1)

AR(2)

MA(1)

MA(2)

ARMA(1,1)

Estimating the ARIMA Model

When you have identified the model as AR, MA, or a mixture, then you can fit it by specifying the AR order (P=*n*) and the MA order (Q=*n*). The "I" in ARIMA stands for "Integrated" and is a parameter that has to do with differencing. Any differencing you do while identifying the model will be included automatically in calculating your forecasts.

When you have estimated the model, pay attention to the standard errors of the parameters. If a parameter estimate is much smaller in absolute value than two standard errors away from zero, then it is probably unnecessary in the model. Refit the model without it. If you are uncertain about model identification, you can sometimes use this rule of thumb to compare two different models. The mean square error (MSE) of the model fit can also guide you. Generally, you are looking for a parsimonious model with small MSE.

Problems with Forecast Models

Forecasting is a vast field, and we cannot begin to explain even the basics in such a brief discussion. Makridakis et al. (1997) cover the topic fairly extensively. SYSTAT contains several methods for forecasting, with which you can experiment on these data. Exponential smoothing, for example, should provide similar forecasts to the ARIMA model. Keep in mind several things as you go:

- There is nothing like extrinsic knowledge. We use forecasting methods for SYSTAT budget planning. We always compare them to staff predictions of sales, however. In general, averaging staff predictions does better than the data-driven forecasting models. The reason is simple-staff know about external factors that are likely to affect sales. These are one-time events that are not easily included in models. Although we are not experts on the stock market, we would bet the same is true for investing. "Chartist" models that are based solely on the trends in stocks will not do as well, on the average, as strategies based on knowledge of companies' economic performance and, in the illegal extreme, inside trading information.
- Always examine your residuals. The same reasons for using residual diagnostics in ordinary linear regression apply to nonlinear forecasting models. In both cases, you want to see independence, or white noise.
- Do not extrapolate too far. As in regression, predictions beyond the data are shaky. The farther you stray from the ends of the data, the less reliable are the predictions. The confidence limits on the forecasts will give you some flavor of this.

Box et al. (1994) provide the primary reference for these procedures. Financial forecasters should consult Nelson (1973) and Vandaele (1983) for applied introductions. Social scientists should look at McCleary and Hay (1980) for applications to behavioral data.

Many treatments (including Box and Jenkins) outline the ARIMA modeling process in three stages: Identification, Estimation, and Diagnosis. This is the outline we have followed in this introduction. With SYSTAT you identify models with Transform, Case plot, ACF plot, and PACF plot, estimate them with ARIMA, and diagnose their adequacy with more plots. For more complex problems, you may also have to use other procedures.

ARIMA models can fit many time series with remarkably few parameters. Sometimes, ARIMA and Fourier models can be used effectively on the same data. As with other modeling procedures, decisions about appropriateness of competing models must rest on theoretical grounds. Nevertheless, a researcher should lean toward ARIMA models when it is reasonable to assume that points in a process are primarily functions of previous points and their errors, rather than periodic signal plus noise.

Seasonal Decomposition and Adjustment

A time series can be viewed as a sum of individual components that may include a term for location (level or mean value), a trend component (long-term movements in the level of a series over time), a seasonal component, and an irregular component (the part unique to each time point). We can use the Mean transformation to remove the mean (location) from a series, Trend to remove a linear trend from a series, and Difference to eliminate either a trend or a seasonal effect from a series. Each of these transformations changes the scale of the series but does not directly provide information about the form of the trend or the seasonal component.

Alternatively, you may want to adjust the values in a series for the seasonal component but leave the series in the same scale or unit. This enables you to interpret the value units in the same way as the original series and to compare values in the series after removing differences due to seasonality.

For example, sales data for many products are strongly seasonal. More suntan lotion is sold in the summer than in the winter. It is therefore difficult to compare suntan lotion sales from month to month (going up? going down?) without first taking seasonal differences into account.

Seasonal differences can be accounted for by determining a factor for each period of the cycle. Quarterly data may have a seasonal factor for each of the four quarters. Monthly data may have a seasonal factor for each of the twelve months.

Seasonal factors can take either of two forms: additive (fixed) or multiplicative (proportional). An additive seasonal factor is a fixed number of units above or below the general level of the series; for example, 10,000 more bottles of suntan lotion were sold in July than the average month. In a multiplicative or proportional model, the seasonal factor is a percentage of the level of the series; for example, 200% more bottles of suntan lotion were sold in July than in the average month.

Additive seasonal effects are removed from a series by subtracting estimates of the appropriate seasonal factor from each point in the series. Multiplicative seasonal effects are removed by dividing each point by the appropriate seasonal factor. Seasonal Adjustment computes either additive or multiplicative seasonal factors for a series and uses them to adjust the original series.

Exponential Smoothing

Exponential smoothing forecasts future observations as weighted averages (a running smoother) of previous observations. For simple exponential smoothing, each forecast is the new estimate of location for the series. For models with trend and/or seasonal components, exponential smoothing smooths the location, trend, and seasonal components separately. For each component, you must specify a smoothing weight between 0 and 1. In practice, weights between 0.10 and 0.30 are most frequently used.

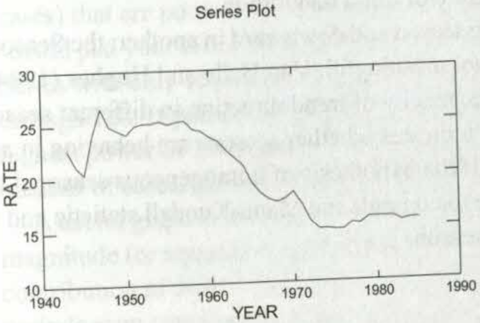
The Exponential Smoothing option allows you to specify a linear or percentage growth (also called exponential or multiplicative) trend or neither, and an additive or multiplicative seasonal component or neither. There is always a location component. Thus, there are nine possible smoothing models from which you can choose.

Smoothing with a linear trend component and no seasonal component is **Holt's method**. Smoothing with both a linear trend and a multiplicative seasonal term is **Winter's three-parameter model**.

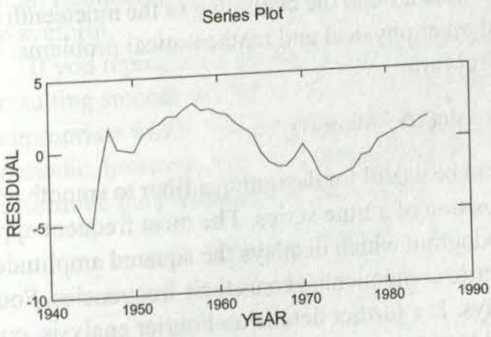
The exponential smoothing procedure obtains initial estimates of seasonal components in the same manner as Seasonal. If there is a trend component, SYSTAT uses regression (after adjusting values for any seasonal effects) to estimate the initial values of the location and trend parameters. If there is neither a trend nor a seasonal component, the first value in the series is used as the initial estimate of location.

Trend Analysis

Consider the example of U.S. birth rates used earlier to explain ARIMA modeling and Forecasting. We are interested in testing the significance of the trend in the birth rate series. Let us look at the time series plot.



It gives some evidence for the presence of a downward trend in the series. We can fit the ordinary least-squares regression line to the data and test for the significance of the slope. Intervention Analysis and Box-Jenkins models can also be used for detecting and estimating the trends (Gilbert, 1987). But these types of parametric techniques are valid only under certain assumptions. In the case of least-squares regression, estimates of the parameters depend on the assumption of independence of residuals. Look at the plot of residuals series from the least-squares regression of birth rate on year. Residuals are clearly not independent.



Often nonparametric trend tests are used to test the significance of the trend in a series and they do not require the assumption of distribution; they also allow missing values.

The Mann-Kendall test is a nonparametric rank based trend test for nonseasonal data. Mann (1945) first suggested using the test for the significance of Kendall's tau where time is considered as one of the variables. Sen (1968) developed a simple and robust estimator of the regression coefficient based on Kendall's rank correlation tau. Hirsch et al. (1982) presented a modification to the Mann-Kendall test, viz., Seasonal Kendall test, for situations where seasonal cycles are present in the series. You can also request for the slope estimator, a measure of trend magnitude.

If the trend is upward in one season and downward in another, the Seasonal Kendall test and slope estimators are not meaningful. Van Belle and Hughes (1984) proposed a test for determining the homogeneity of trend direction in different seasons. This provides a single statistic that indicates whether seasons are behaving in a homogeneous manner or not. If the hypothesis of homogeneous seasonal trends over time is rejected, it is advisable to compute the Mann-Kendall statistic and the slope estimator for each individual season.

Fourier Analysis

If you believe your series is cyclical — such as astronomical or behavioral data — then you should consider **Fourier analysis**. The Fourier model decomposes a series into a finite sum of trigonometric components — sine and cosine waves of different frequencies. If your data are cyclical at a particular frequency, such as monthly, then a few Fourier components might be sufficient to capture most of the nonrandom variation.

Fourier analysis decomposes a time series just as a musical waveform can be decomposed into a fundamental wave plus harmonics. The French mathematician Fourier devised this decomposition around the beginning of the nineteenth century and applied it to heat transfer and other physical and mathematical problems. This transformation is of the general form:

$$f(t) = x + x\sin(t) + x\cos(t) + x\sin(2t) + x\cos(2t) + \ldots$$

The Fourier decomposition can be useful for designing a filter to smooth noise and for analyzing the spectral composition of a time series. The most frequent application involves constructing a periodogram which displays the squared amplitude (magnitude) of the trigonometric components versus their frequencies. Fourier can be used to construct these displays. For further details on Fourier analysis, consult Brigham (1988) or Bloomfield (2000).

Fourier transforms are time consuming to compute because they involve numerous trigonometric functions. Cooley and Tukey (1965) developed a fast algorithm for computing the transform on a discrete series that makes the spectral analysis of lengthy series practical. A variant of this Fast Fourier Transform algorithm is implemented in SYSTAT.

The discrete Fourier transform should be done on series with lengths (number of cases) that are powers of 2. If you do not have samples of 32, 64, 128, 256, etc., you should pad your series with zeros up to the next power of 2. If you have a series called Series with only 102 cases, for example, you can recode to add zeros to cases 103 through 128. If you do not pad the file in this way, the Fourier procedure finds the highest power of 2 less than the number of cases in the file and transforms only that number of cases. (In this example, it would have transformed only the first 64 cases.)

A useful graph to accompany Fourier analysis is the **periodogram**. This graph plots magnitude (or squared magnitude) against frequency. It reveals the relative contribution of different frequency waveforms to the overall shape of the series. If the periodogram contains one large spike, then it means that the series can be fit well by a single sinusoidal waveform. The periodogram is itself like a series, so sometimes you may want to smooth it with one of the Series smoothers.

Fourier analysis is often used to construct a filter, which works like running smoothers. A filter allows variation of only a limited band of frequencies to pass through. A low-pass filter, for example, removes high-frequency information. It is often used to remove noise in radio transmissions, recordings, and photographs. A high-pass filter, on the other hand, removes low-frequency variation. It is used as one method for detecting edges in photographs. You can construct filters in SYSTAT by computing the Fourier transform, deleting real and imaginary components for low or high frequencies, and then using the inverse transform to produce a smoothed waveform.

If you reproduce a series from a few low-frequency Fourier components, the resulting smooth will be similar to that achieved by a running window of an appropriate width. The Fourier method will constrain the smooth to be more regularly periodic, however, since the selected trigonometric components will completely determine the periodicity of the smooth.

Graphical Displays for Time Series in SYSTAT

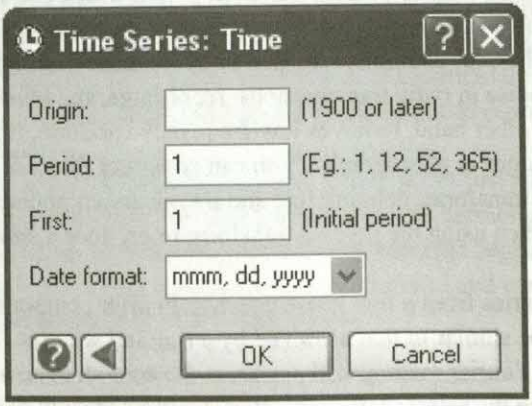
Plotting data, autocorrelations, and partial autocorrelations is often one of the first steps in understanding time series data. SYSTAT provides several graphical displays, each of which is discussed in turn.

Time Axis Format Dialog Box

Time Axis Format labels the sequence(s) of values in a file with identifiers that represent the cycle and the periodicity. The identifiers label the time series plot *x*-axis for each time point.

To open the Time Axis Format dialog box, from the menus choose:

Analyze
  Time Series
    Time...



The following options can be specified:

Origin. Starting point of the time series, expressed as a year.

Period. Periods within each year. The value defines the number of observations within each year. For example, specify 12 for months, 52 for weeks, 365 for days, etc. If there is only one observation per year, specify 1.

First. Starting point of the period of observation. For example, if the period is months within each year and the first observation is for June, the First value would be 6.

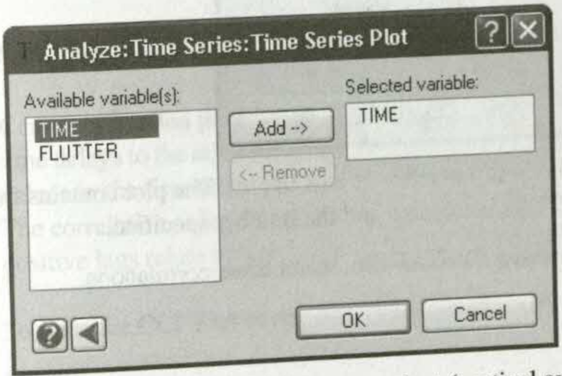
Date format. The date display format for values on the *x*-axis (time axis). Select a format from the drop-down list.

Time Series Plot Dialog Box

Time series plot can give you a general idea of a series, enabling you to detect a long-term trend, seasonal fluctuations, and gross outliers.

To open the Time Series Plot dialog box, from the menus choose:

Analyze
  Time Series
    Time Series Plot...



The variable you select is the dependent (vertical axis) variable, and the case number (time series observation) is the independent variable (horizontal axis). The points are connected with a line.

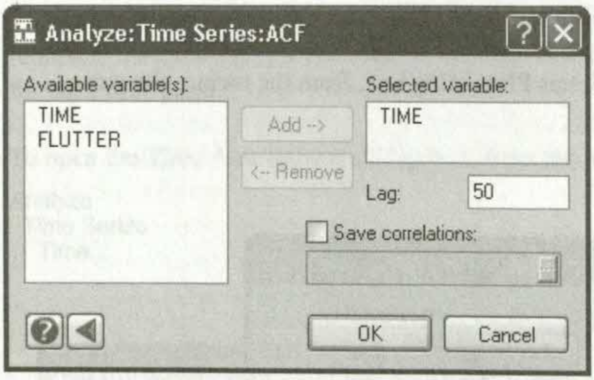
ACF Plot Dialog Box

Autocorrelation plots show the correlations of a time series with itself shifted down a specified number of cases. Plots of autocorrelations help you to investigate the relation of each time point to previous time points. If the autocorrelation at lag 1 is high, then each value is highly correlated with the value at the previous time point. If the autocorrelation at lag 12 is high for data collected monthly, then each month is highly

correlated with the same month a year before (for example, for monthly sales data, sales in December may be more related to those in previous December's than to those in November or January).

To open the ACF Plot dialog box, from the menus choose:

Analyze
 Time Series
  ACF...



You can specify the maximum number of lags to plot. The plot contains the autocorrelations for all lags between 1 and the number specified.

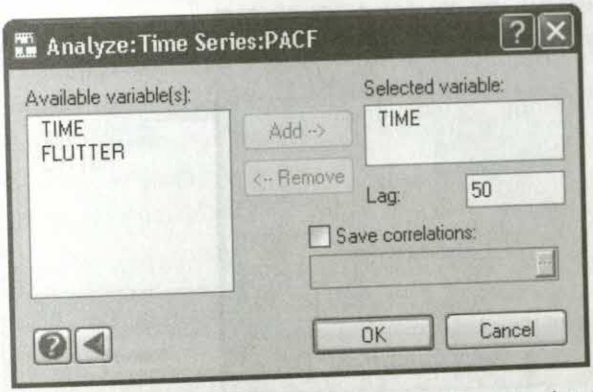
To save the correlations to a SYSTAT file, select Save correlations.

PACF Plot Dialog Box

Partial autocorrelation plots show the relationship of points in a series to preceding points after partialing out the influence of intervening points.

To open the PACF Plot dialog box, from the menus choose:

Analyze
 Time Series
  PACF...



You can specify the maximum number of lags to plot. The plot contains the partial autocorrelations for all lags between 1 and the number specified.

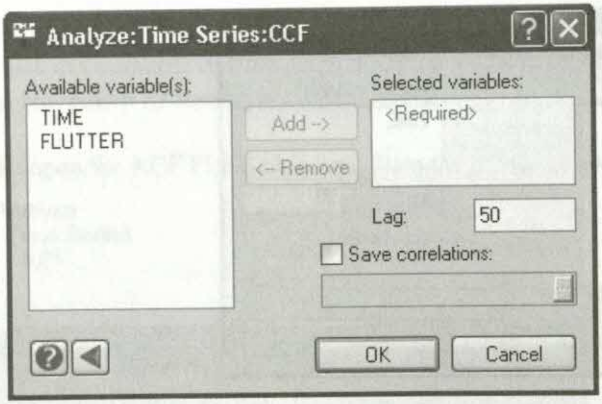
To save the correlations to a SYSTAT file, select Save correlations.

CCF Plot Dialog Box

Cross-correlation plots help to identify relations between two different series and any time delays to the relations. A correlation for a negative lag indicates the relation of the values in the first series to values in the second series that number of periods earlier. The correlation at lag 0 is the usual Pearson correlation. Similarly, correlations at positive lags relate values in the first series to subsequent values in the second series.

To open the CCF Plot dialog box, from the menus choose:

Analyze
  Time Series
    CCF...



You can specify the number of lags to plot. Approximately half of the lags will be positive and half will be negative.

To save the correlations to a SYSTAT file, select Save correlations.

Using Commands

To graph a time series, first specify your data with USE *filename*. Continue with:

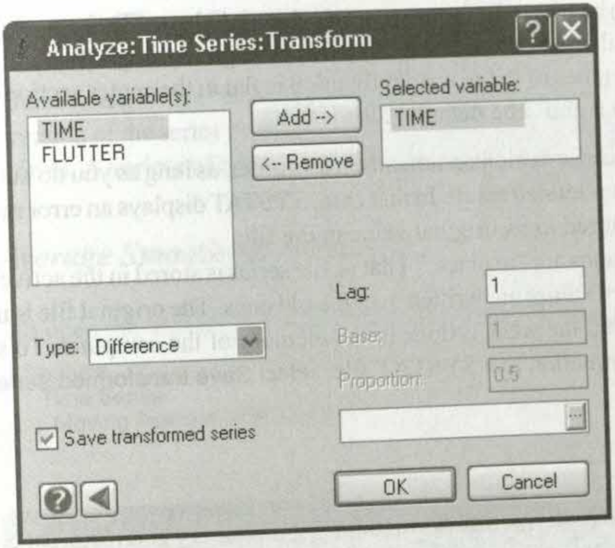
```
SERIES
      TIME origin period first
      TPLOT var / LAG=n
      ACF var / LAG=n
      PACF var / LAG=n
      CCF var1, var2 / LAG=n
```

Transformations of Time Series in SYSTAT

Transform Dialog Box

To open the Transform dialog box, from the menus choose:

Analyze
  Time Series
    Transform...



Available transformations include:

- **Mean.** Subtracts the mean from each value in the series.
- **Log**. Replaces the values in a series with their natural logarithms, and thus removes nonstationary variability, such as increasing variability over time.
- **Square**. Squares the values in a series. This is useful for producing periodograms and for normalizing variance across the series.
- **Trend**. Removes linear trend from a series.
- **Difference**. Replaces each value by the difference between it and the previous value, thereby removing trend (nonstationarity in level over time). Using differences between each successive value is called lag 1. A Lag option allows seasonal differences (for example, for data collected monthly, request a lag of 12).
- **Percent change**. Replaces each value by the difference from the previous value expressed as a percentage change-the difference in values divided by the previous value.
- **Index**. Replaces each value by the ratio of the value to the value of a base observation, which you can specify for Base. By default, SYSTAT uses the first observation in the series.
- **Taper**. Smooths the series with the split-cosine-bell taper. Tapering weights the middle of a series more than the endpoints. Use it prior to a Fourier decomposition to reduce "leakage" between components. For Proportion, enter the proportion (P)

of the series to be tapered. Choose a weight function that varies between a "boxcar" (P=0) and a full cycle of a cosine wave from trough to trough (P=1). For intermediate values of P, the weight function is flat in the center section and cosine tapered at either end. The default value is 0.5.

You can pile up transformation commands in any order, as long as you do not encounter a mathematically undefined result. In that case, SYSTAT displays an error message and the variable is restored to its original value in the file.

All transformations are "in place." That is, the series is stored in the active work area and the transformed values are written over the old ones. The original file is not altered, however, because all the work is done in the memory of the computer. To save the results of a transformation to a SYSTAT file, select Save transformed series.

Clear Series

You can clear any past series transformations from memory and restore the original values of the series. It is not possible to clear only the latest transformation (unless you are saving to files after each step). Clear Series undoes all the transformations.

To clear series transformations, from the menus choose:

Analyze
  Time Series
    Clear Series...

Using Commands

To transform a time series, first specify your data with USE *filename*. Continue with:

```
SERIES
       DIFFERENCE var / LAG=n
       LOG var
       PCNTCHANGE var
       MEAN var
       SQUARE var
       TREND var
       INDEX var / BASE=n
       TAPER var / P=n
```

CLEAR var clears transformations from memory.

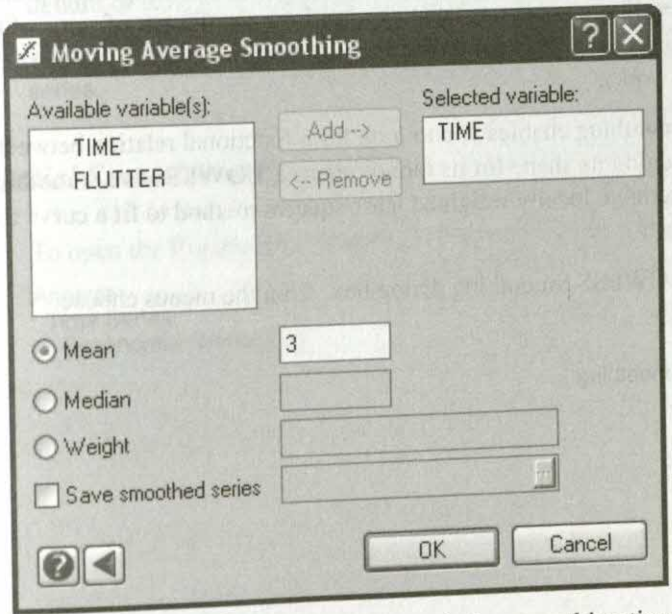
Smoothing a Time Series in SYSTAT

Sometimes, with a "noisy" time series, you simply want to view some sort of smoothed version of the series even though you have no idea what type of function generated the series. A variety of techniques can smooth, or filter, the noise from such a series.

Moving Average Smoothing Dialog Box

To open the Moving Average Smoothing dialog box, from the menus choose:

Analyze
  Time Series
    Moving Average Smoothing...



Smooth provides the following methods for smoothing time series:

- **Mean**. Running means (moving averages). Mean of a span of series values surrounding and including the current value. Specify the number of values (observations) to use in the calculation. Default value is 3.

- **Median**. Running medians. Median of a span of series values surrounding and including the current value. Specify the number of values (observations) to use in the calculation.
- **Weight**. General linear filters in which you can specify your own weights. Smooth transforms the weights before using them so that they sum to 1.0. Weight=1, 2, 1 is the same as Weight = 0.25, 0.5, 0.25 or Weight = 3, 6, 3.

To save the results of a smoothing operation to a SYSTAT file, select Save smoothed series.

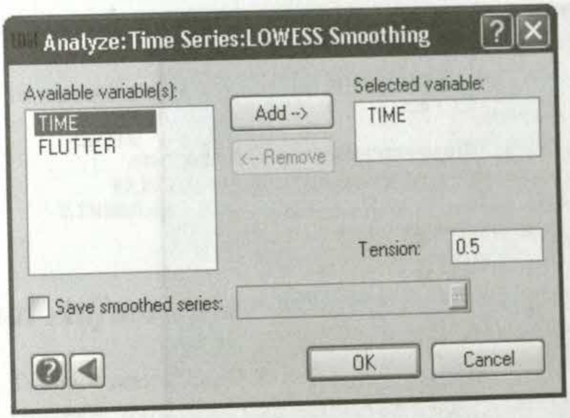
LOWESS Smoothing Dialog Box

Cleveland (1979) presented a method for smoothing values of *Y* paired with a set of ordered *X* values. Chambers et al. (1983) introduce this technique and present some clear examples. If you are not a statistician, and want a glimpse of some of the details, read Chambers et al. (1983) and Velleman and Hoaglin (1981) (if you are unfamiliar with Tukey's work).

Scatterplot smoothing enables you to look for a functional relation between *Y* and *X* without prejudging its shape (or its monotonicity). **LOWESS** is a smoothing method that uses an iterative locally weighted least-squares method to fit a curve to a set of points.

To open the LOWESS Smoothing dialog box, from the menus choose:

Analyze
  Time Series
    LOWESS Smoothing...



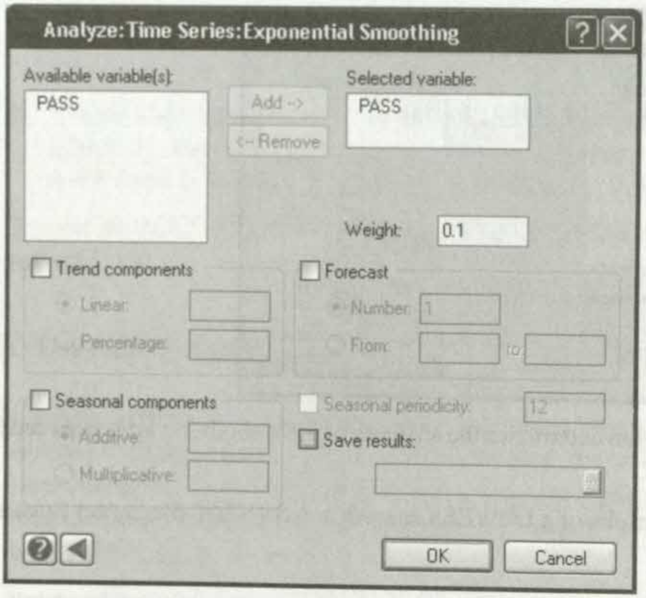
Tension. Tension determines the stiffness of the smooth. It varies between 0 and 1 with default of 0.5.

To save the results of a LOWESS smooth to a SYSTAT file, select Save smoothed series.

Exponential Smoothing Dialog Box

To open the Exponential Smoothing dialog box, from the menus choose:

Analyze
  Time Series
    Exponential Smoothing...



The following options can be specified:

Weight. Specify a smoothing weight between 0 and 1. In practice, weights between 0.1 and 0.3 are most frequently used. The default value is 0.1.

Trend components. You can supply a weight for either a Linear or Percentage trend component. Values usually range between 0.1 and 0.3.

Forecast. Number or the range of new cases to predict. For example, a value of 10 produces forecasts for 10 time points; a range from 144 to 154 produces forecasts for time points 144 through 154. By default, the forecast for one time point beyond the range of data set, is given.

Seasonal components. You can supply a weight for either Additive or Multiplicative. Values usually range between 0.1 and 0.3.

Seasonal periodicity. Indicates the repetitive cyclical variation, such as the number of months in a year or the number of days in a week. The default value is 12 (as in months in a year) unless the period is already specified in Time Axis Format dialog box.

To save the forecasts and residuals to a SYSTAT file, select Save results.

Using Commands

To smooth a time series, first specify your data with USE *filename*. Continue with:

```
SERIES
      SMOOTH var / LOWESS=n MEAN=n  MEDIAN=n  WT=n1,n2,…
    EXPONENTIAL var / ADDITIVE=n  FORECAST=n (or a,b for a range)
LINEAR=n  MULTIPLICATIVE=n PERCENTAGE=n, SEASON=n  SMOOTH=n
```

Seasonal Adjustments in SYSTAT

Transformations can remove the mean, trends, and seasonal effects from a time series. However, transformations alter the scale of a time series and also yield no information regarding the form of the removed trend or seasonal effect. As an alternative, you can use seasonal adjustments to account for seasonal factors while maintaining the original scale of the time series.

Seasonal Adjustment Dialog Box

To open the Seasonal Adjustment dialog box, from the menus choose:

Analyze
  Time Series
    Seasonal Adjustment...

The following options can be specified:

- **Term**. An Additive seasonal factor is a fixed number of units above or below the general level of the series. In a Multiplicative model, the seasonal factor is a percentage of the level of the series.
- **Season**. Indicates the **periodicity** — the repetitive cyclical variation — such as the number of months in a year or the number of days in a week. The default value is 12.

To save the deseasonalized series to a SYSTAT file, select Save adjusted series.

Using Commands

To seasonally adjust a time series, first specify your data with USE *filename*. Continue with:

```
SERIES
        ADJSEASON var / ADDITIVE  MULTIPLICATIVE  SEASON=n
```

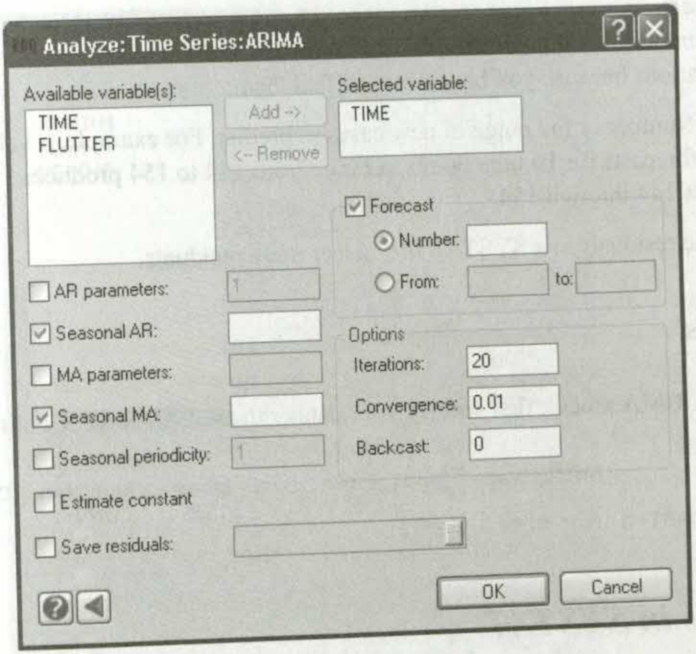
ARIMA Models in SYSTAT

ARIMA (AutoRegressive Integrated Moving Average) models combine autoregressive techniques with the moving average approach. Consequently, each case is a function of previous cases and previous errors.

ARIMA Dialog Box

To open the ARIMA dialog box, from the menus choose:

Analyze
  Time Series
    ARIMA...



The following options can be specified:

AR parameters. Number of autoregressive parameters. The default value is 1.

Seasonal AR. Number of seasonal autoregressive parameters.

MA parameter. Number of moving average parameters.

Seasonal MA. Number of seasonal moving average parameters.

Seasonal periodicity. Defines the seasonal periodicity. The default value is 1.

Estimate constant. Includes a constant in the model.

Options. You can specify the number of iterations (Default value is 20) for the ARIMA model, the convergence criterion, (Default value is 0.01) and a backcast (Default value is 0) value to extend the series backwards (forecasting in reverse). Although it slows down computation, you should use backcasting for seasonal models especially and choose a length greater than the seasonal period.

Do not play around with convergence unless you are failing to get convergence of the estimates after many iterations. It is better to increase the number of iterations than

to decrease the convergence criterion, since your estimates will be more precise. In any case, it cannot be set greater than one tenth. Sometimes models fail to converge after many iterations because you have misspecified them.

Forecast. Number or the range of new cases to predict. For example, a value of 10 produces forecasts for 10 time points; a range from 144 to 154 produces forecasts for time points 144 through 154.

To save the residuals to a SYSTAT file, select Save residuals.

Using Commands

To fit an ARIMA model, first specify your data with USE *filename*. Continue with:

```
SERIES
                ARIMA var / P=n  PS=n  Q=n  QS=n  SEASON=n  CONSTANT,
                                    BACKCAST=n  ITER=n  CONV=n,
FORECAST=n (or time1,time2)
```

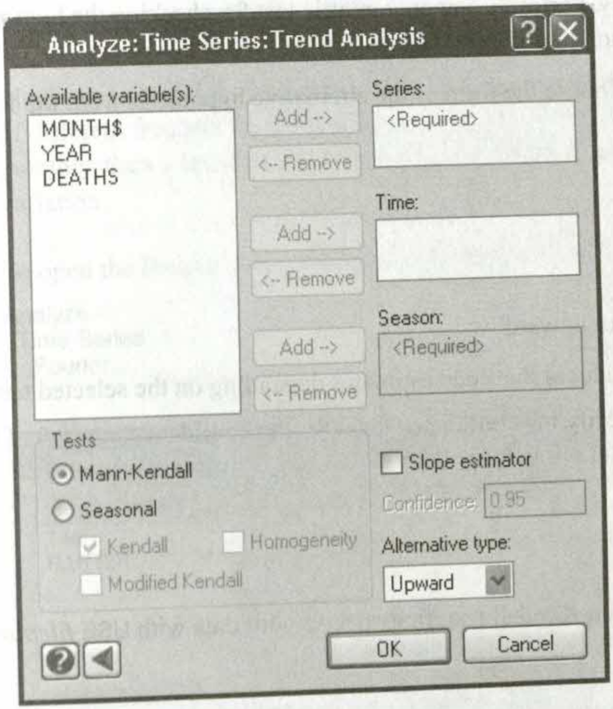
Trend Analysis in SYSTAT

Trend analysis provides nonparametric trend tests for detecting and estimating the trends in a given time series data.

Trend Analysis dialog box

To open the Trend Analysis dialog box, from the menus choose:

Analyze
 Time Series
  Trend Analysis…



Series. Select the series variable

Time. Select the time variable

Season. Select the seasonal variable; the seasonal variable reports the season in which the observations are collected. The observations collected in one season should have the same name or value as for the seasonal variable. The default value is 12.

Tests. Select the Mann-Kendall test or any of the different types of Seasonal tests from the following options:

- **Mann-Kendall.** Performs a simple (nonparametric) rank-based test for assessing the significance of the trend in a time series.
- **Seasonal.** The following tests are available to detect seasonal trends in the data:
  - **Kendall.** Performs a nonparametric trend test for seasonal data by assuming independence among seasons.
  - **Modified Kendall.** Performs a nonparametric trend test for seasonal data by considering the covariance among seasons.

- **Homogeneity.** Performs a nonparametric test for checking the homogeneity of trend direction in different seasons.

Alternative type. Choose the form of the alternative hypothesis from the following three options:

- upward
- downward
- twosided

The default option is 'upward'.

Slope estimator. Produces the slope estimator depending on the selected test.

- **Confidence.** Specify the confidence level for the confidence interval of the slope estimator. The default value is 0.95.

Using Commands

To carry out the Mann-Kendall test, first specify your data with USE *filename*. Continue with:

```
SERIES
MKTEST SERIES*TIME / ALT=ALTER SLOPE CONFI=U
```

To carry out the Seasonal trend test(s), first specify your data with USE *filename*. Continue with:

```
SERIES
STEST SERIES*TIME / TEST=TEST SEASON=SEASON ALT=ALTER SLOPE
CONFI=U
```

TEST may be SK or MSK or HT and
ALTER may be UPWARD or DOWNWARD or TWOSIDED.
UPWARD is the default option.

Fourier Models in SYSTAT

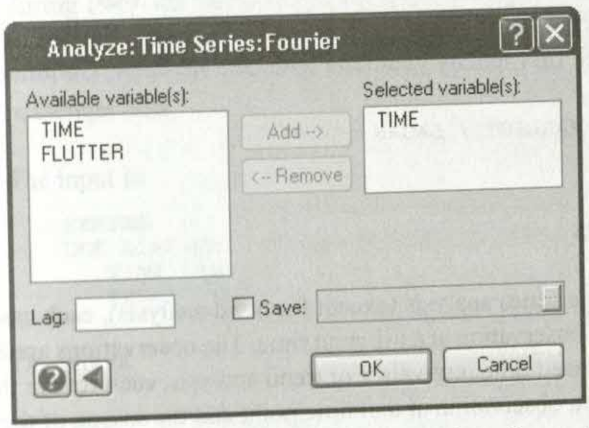
Fourier models are particularly well-suited to cyclical time series. These models decompose a time series into a sum of trigonometric components.

Fourier Transformation Dialog Box

The Fourier model decomposes a time series into a finite sum of sine and cosine waves of different frequencies. If your data are cyclical at a particular frequency, such as monthly, then a few Fourier components might capture most of the nonrandom variation.

To open the Fourier Transformation dialog box, from the menus choose:

Analyze
  Time Series
    Fourier...



The Lag specification indicates the number of cases to use in the analysis.

If you select two variables, the inverse transformation is computed. The first variable selected is used as the real component, and the second variable is used as the imaginary component. To save the real and imaginary components in a SYSTAT file, select Save. Real and imaginary components are saved instead of magnitude and phase because that allows you to do an inverse Fourier transform.

For example, assume that you have saved the results of a direct transformation into a file *MYFOUR*. That file should contain two variables — *REAL* and *IMAG* — which are the two components of the transformation. To obtain the inverse Fourier transformation:

```
USE MYFOUR
FOURIER REAL IMAG
```

Since you specify two variables, SYSTAT assumes that you want the inverse transformation, and that the first variable is the real component, and the second, the imaginary component. The work is done in the active work area, so the resulting real series is stored in the active work area occupied by *REAL* (or whatever you called the first variable corresponding to the real component).

If you absolutely must have magnitude and phase in a SYSTAT file instead of the real and imaginary components, do the following transformations:

```
USE MYFOUR
LET MAGNITUDE = SQR(REAL*REAL + IMAG*IMAG)
LET PHASE = ATN (IMAG/REAL)
```

Using Commands

To fit a Fourier model, first specify your data with USE *filename*. Continue with:

```
SERIES
      FOURIER VARLIST / LAG=N
```

Usage Considerations

Types of data. For time series analysis (except for trend analysis), each case (row) in the data represents an observation at a different time. The observations are assumed to be taken at equally spaced time intervals. For trend analysis, each row in the data represents a time point, observation at that time point and the season of the measurement.

Print options. Output is standard for all PLENGTH options.

Quick Graphs. Smoothing, seasonal adjustments, and the Mann Kendall test yield a time series plot. Forecasting in ARIMA results in a time series plot of the original series with the forecasts. The Seasonal Kendall, the Modified Seasonal Kendall, and the Homogeneity test yield a time series plot across all seasons. Fourier analysis produces periodograms (the squared magnitude against frequencies).

Saving files. You can save ACF, PACF, CCF, transformed, smoothed, deseasonalized, and forecasted values, as well as both the real and imaginary parts of the Fourier transform.

BY groups. BY groups has no effect in SERIES.

Case frequencies. FREQUENCY variable has no effect in SERIES.

Case weights. SERIES does not allow case weighting.

Examples

Example 1
Time Series Plot

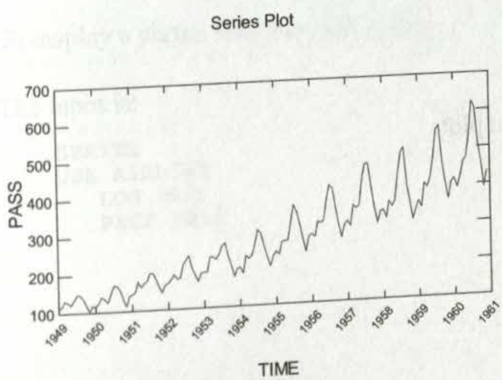
To illustrate these displays, we use monthly counts of international airline passengers during 1949–60. Box et al. (1994) call the series G. Each of the 144 monthly counts is stored as a case in the SYSTAT file named *AIRLINE*.

TPLOT provides a graphical view of the raw data. Here we plot the *AIRLINE* passenger data.

The input is:

```
SERIES
USE AIRLINE
   TIME 1949 12
   TPLOT PASS
```

The plot is:



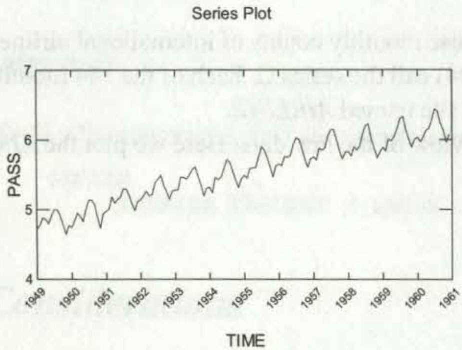
Notice that the counts tend to peak during the summer months each year and that the number of passengers tends to increase over time (a positive trend). Notice also that

the spread or variance tends to increase over time. One way to deal with this problem is to log-transform the data.

Applying the log transformation requires the following commands:

```
LOG PASS
TPLOT PASS
```

The plot is:



Compare this plot with the previous one — the variance across time now appears more stable, but there is still a positive upward trend over time.

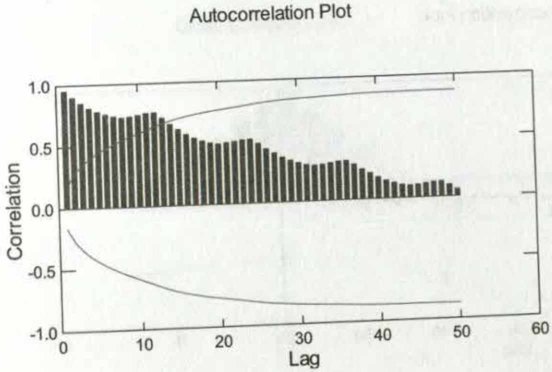
Example 2
Autocorrelation Plot

To display an autocorrelation plot:

The input is:

```
SERIES
USE AIRLINE
   LOG PASS
   ACF PASS
```

The plot is:



Note that we use the logged values of *PASS*. The shading in the display indicates the size of the correlation at each lag (that is, like a bar chart). The correlation of each value with the previous value in time (lag 1) is close to 1.0; with values 12 months before (lag 12), it is around 0.75. The curved line marks approximate 95% confidence levels for the significance of each correlation. Notice the slow decay of these values. To most investigators, this indicates that the series should be differenced.

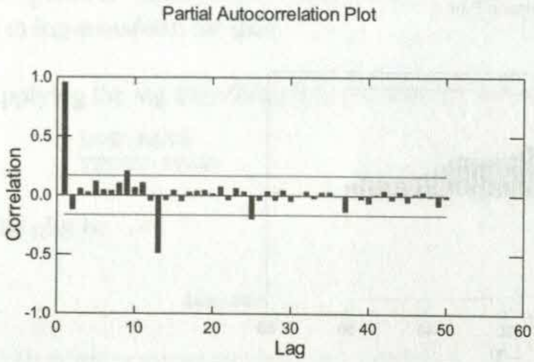
Example 3
Partial Autocorrelation Plot

To display a partial autocorrelation plot:

The input is:

```
SERIES
USE AIRLINE
    LOG PASS
    PACF PASS
```

The plot is:



The first autocorrelation is the same as in the ACF plot. There are no previous autocorrelations, so it is not adjusted. The second-order autocorrelation was close to 0.90 in the ACF plot, but after adjusting for the first autocorrelation, it is reduced to −0.118.

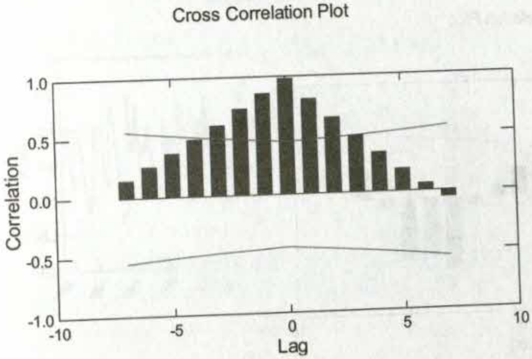
Example 4
Cross-Correlation Plot

This example uses the *SPNDMONY* file, which contains two quarterly series, *SPENDING* (consumer expenditures) and *MONEY* (money stock) in billions of current dollars for the United States during the years 1952–1956. The first record (case) in the file contains the *SPENDING* and *MONEY* dollars for the first quarter of 1952; the second record, dollars for the second quarter of 1952, and so on (that is, if each case contains *SPENDING* and *MONEY* values for a quarter). These series are analyzed by Chatterjee and Price (2006).

The input is:

```
SERIES
USE SPNDMONY
   CCF SPENDING MONEY / LAG=15
```

The plot is:



There is strong correlation between the two series at lag 0, tapering off the further one goes in either direction. This is true of all cross-correlation functions between two trended series. Since both series are increasing, early values in both series tend to be small, and final values tend to be large. This produces a large positive correlation.

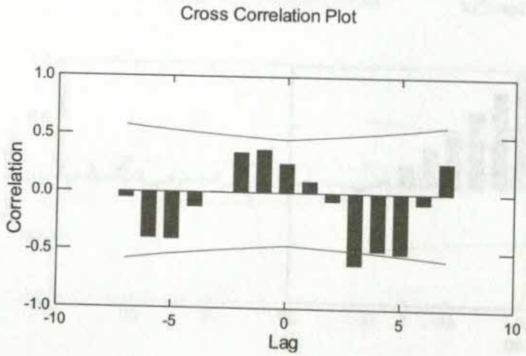
Differencing

To better understand the relationship, if any, between the series, difference them to remove the common trend and then display a new CCF plot.

The input is:

```
SERIES
USE SPNDMONY
   DIFFERENCE SPENDING
   DIFFERENCE MONEY
   CCF SPENDING MONEY / LAG=15
```

The plot is:



This shows a significant negative correlation at only one time interval: +3 lags of the series. Since we selected *SPENDING* first, we see that consumer expenditures are negatively correlated with the money stock three quarters later. Thus, consumer spending may be a "leading indicator" of money stock.

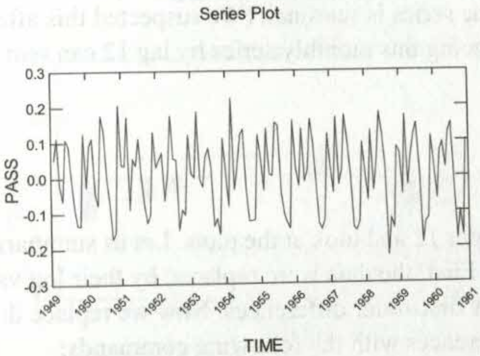
Example 5
Differencing

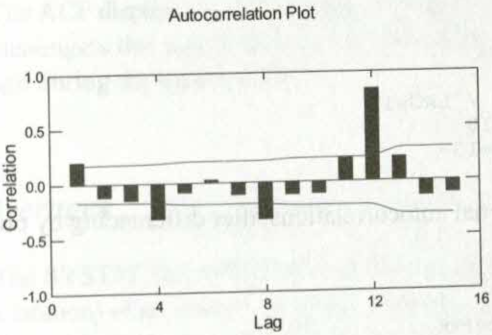
Let us replace the values of the series in the *AIRLINE* data with the difference between each value and the previous value-first order (lag) differencing.

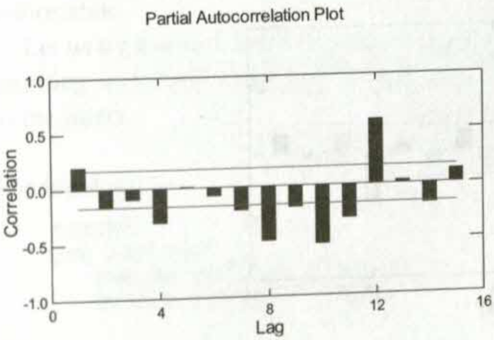
The input is:

```
SERIES
USE AIRLINE
  TIME 1949 12
  LOG PASS
  DIFFERENCE PASS
  TPLOT PASS
  ACF PASS / LAG=15
  PACF PASS / LAG=15
```

The plots are:







The strong upward trend seen in the undifferentiated time series plots is not evident here. Notice also that the scale on this plot ranges from approximately –0.2 to +0.2, while on the plot in the time series plot example, it ranges from 4.6 to 6.4.

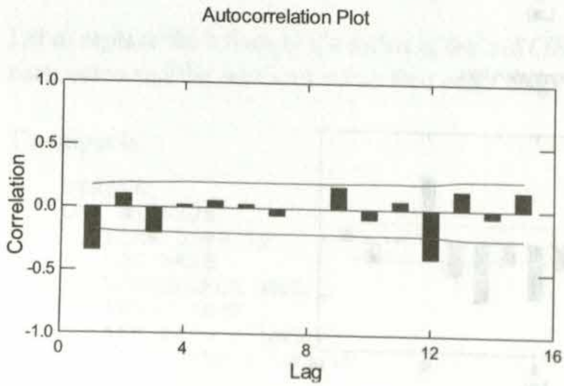
The very strong lag 12 ACF and PACF correlations with a decay of strong correlations for shorter lags suggest that the series is seasonal. (We suspected this after seeing the first plot of the data.) Differencing this monthly series by lag 12 can remove cycles from the series.

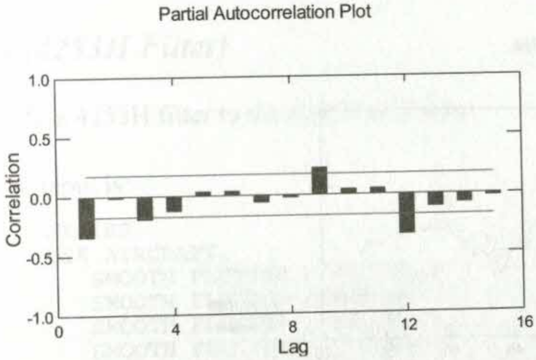
Order 12 Differencing

Here, we will difference by order 12 and look at the plots. Let us summarize what has happened to the original data. First, the data were replaced by their log values. Next, the data were replaced by their first-order differences. Now we replace these differences with order 12 differences with the following commands:

```
SERIES
USE AIRLINE
   LOG PASS
   DIFFERENCE PASS
   DIFFERENCE PASS / LAG=12
   ACF PASS / LAG=15
   PACF PASS / LAG=15
```

The autocorrelations and partial autocorrelations after differencing by order 12 are shown below:





The ACF display has spikes at lag 1 and lag 12. We conclude that the number of airline passengers this month depends on the number last month and on the number one year ago during the same month.

Example 6
Moving Averages

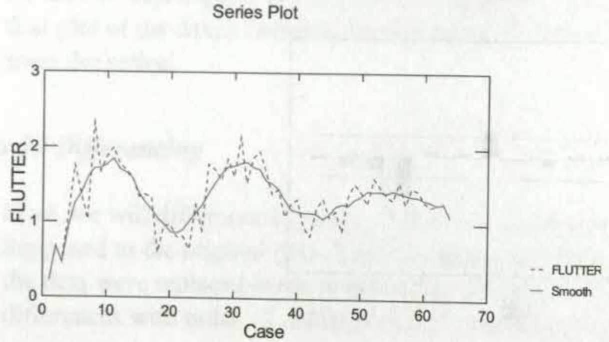
The SYSTAT file *AIRCRAFT* contains the results of a "flutter test" (amplitude of vibration) of an aircraft wing (Bennett and Desmarais, 1975). Although the model for these data is known, we are going to try to recover a smooth series without using this information.

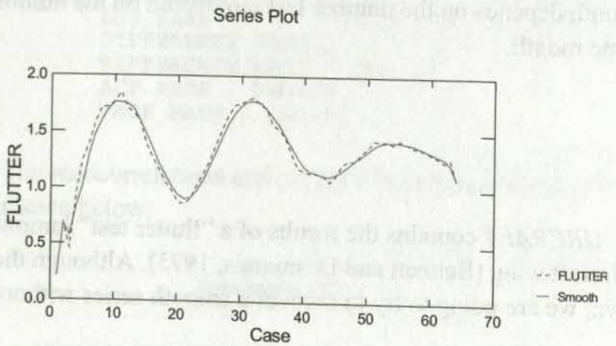
Let us try a seven-point moving average on the *AIRCRAFT* data and smooth the resulting smoothed series with a four-point moving average. This should remove some of the jitters.

The input is:

```
SERIES
USE AIRCRAFT
   SMOOTH FLUTTER / MEAN=7
   SMOOTH FLUTTER / MEAN=4
```

The plots are:





The second plot is even smoother than the first. We chose the lengths of the window by trial and error after looking at the data to see how much they "jitter" to the left and right of each point relative to the overall pattern of the series. You will do better if you know something about the function generating the data.

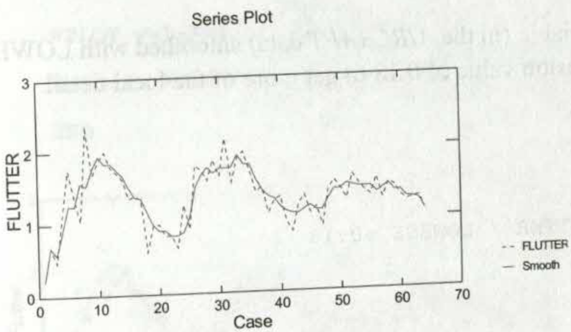
Example 7
Smoothing (4253H Filter)

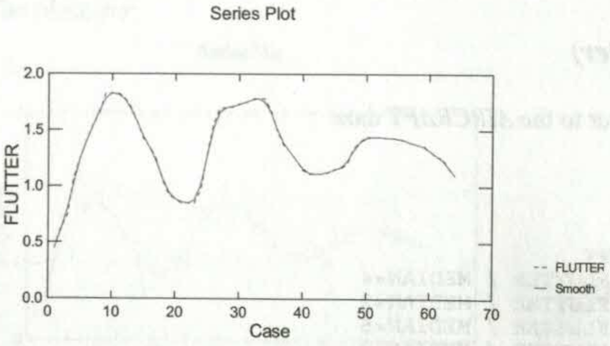
To fit a 4253H filter to the *AIRCRAFT* data:

The input is:

```
SERIES
USE AIRCRAFT
    SMOOTH FLUTTER / MEDIAN=4
    SMOOTH FLUTTER / MEDIAN=2
    SMOOTH FLUTTER / MEDIAN=5
    SMOOTH FLUTTER / MEDIAN=3
    SMOOTH FLUTTER / WT=1,2,1
```

A Quick Graph follows each Smooth request. (To omit the display, type GRAPH=NONE.) The displays shown below correspond to the first request (MEDIAN=4) and the final smooth (a running means smoother with weights).





The previous smooth (MEDIAN=3) is marked by dashed lines.

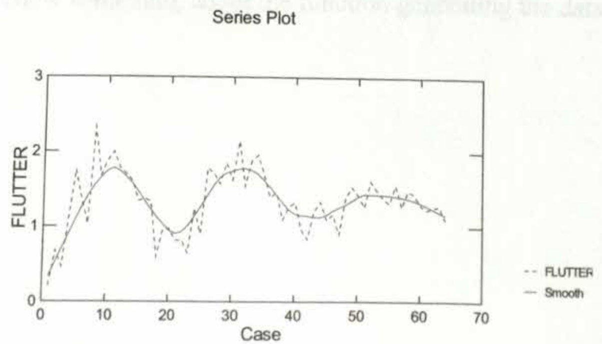
Example 8
LOWESS Smoothing

Here is the flutter variable (in the *AIRCRAFT* data) smoothed with LOWESS smoothing. Use a Tension value of 0.18 to get more of the local detail.

The input is:

```
SERIES
USE AIRCRAFT
   SMOOTH FLUTTER / LOWESS =0.18
```

The plot is:



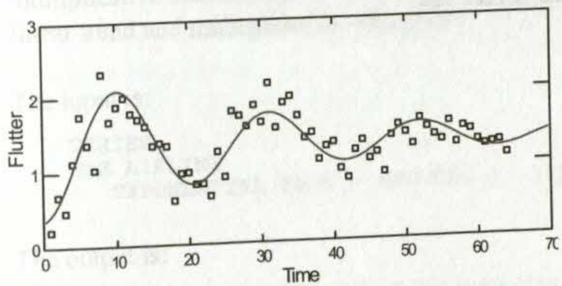
And the Winner Is…

The actual function used to generate the data in the moving average and 4253H filter examples is shown below:

$$Y(t) = 1 - e^{-0.03t}\cos(0.3t)$$

where $t = 1, 2, \ldots, 64$ (the index number of the series). We added normal (Gaussian) noise to this function in inverse proportion to the square root of t. We leave it to the reader to design an optimal filter for the Weight option after looking at the noise distribution in the plot. The generating function on the data is shown below:

```
   USE AIRCRAFT
   BEGIN
PLOT FLUTTER * TIME / HEI=1.5IN  WID=3.5IN,
               XMIN=0  XMAX=70  YMIN=0 YMAX=3,
 XLABEL='Time' YLABEL='Flutter',
               SYMB=1  SIZE=.75  FILL=1  COLOR=BLACK
FPLOT y=1-EXP(-0.03*t)*COS(.3*t) + (0.35) ; ,
               HEI=1.5IN  WID=3.5IN,
XMIN=0  XMAX=70  YMIN=0  YMAX=3,
               XLAB='' YLAB='' AXES=NONE  SCALE=NONE
END
```



Is there a winner? The LOWESS smooth looks pretty good. Usually, for Gaussian data like these, it is hard to beat running means. Running medians and LOWESS do extremely well on non-Gaussian data, however, because they are less susceptible to outliers in the series. You will also find that exploratory smoothing requires a lot of fine tuning with window widths (tension) and weights.

Example 9
Multiplicative Seasonal Factor

We use the same *AIRLINE* data — from Box et al. (1994) — used in the time series plot example. If you examine the plot there, you can see the strong periodicities. The size of the periodicities depends on the level of the series, so we know that the form of seasonality is multiplicative. Each year, the number of passengers peaks during July and August, but there are also jagged spikes in the data that correspond, apparently, to holidays like Christmas and Easter.

Here we adjust the airline series for the multiplicative seasonal effect implied by the series plot.

The input is:

```
SERIES
USE AIRLINE
   TIME 1949 12
   ADJSEASON PASS / MULTIPLICATIVE
```

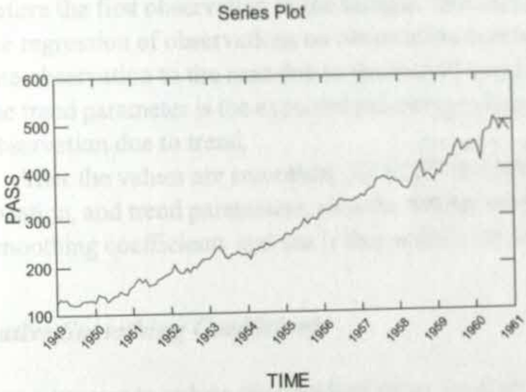
The output is:

```
PASS
Series Originates at: 1949. Periodicity: 12. First Period: 1
PASS copied from SYSTAT file into active work area
Adjust Series for a Seasonal Periodicity of 12
```

Seasonal Indices for the Series are

```
  1     91.077
  2     88.133
  3    100.825
  4     97.321
  5     98.305
  6    111.296
  7    122.636
  8    121.652
  9    105.997
 10     92.200
 11     80.397
 12     90.164
```

Airline travel appears heaviest during the summer months, June (6) through September (9).



The plot shows that the trend and the irregular components remain, but the seasonal component has been removed from the series.

Example 10
Multiplicative Seasonality with a Linear Trend

In the time series plot example, we looked at the *AIRLINE* data from Box et al. (1994). The plot of the series shows a strong increasing trend and what looks like multiplicative seasonality. We could try to forecast this series with a model having a linear trend and multiplicative seasonality.

The input is:

```
SERIES
USE AIRLINE
  EXPONENTIAL PASS / SMOOTH=.3  LINEAR=.4  MULT=.4 FORECAST=10
```

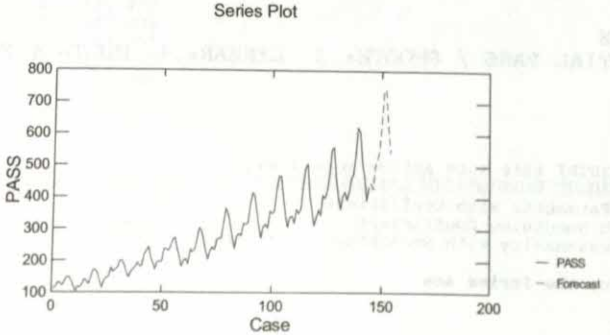
The output is:

```
PASS copied from SYSTAT file into active work area

  Smooth Location Parameter with Coefficient            : 0.300
  Linear Trend with Smoothing Coefficient               : 0.400
  Multiplicative Seasonality with Smoothing Coefficient : 0.400

Seasonal Indices for the Series are

     1     91.077
     2     88.133
     3    100.825
     4     97.321
     5     98.305
     6    111.296
     7    122.636
```

```
   8    121.652
   9    105.997
  10     92.200
  11     80.397
  12     90.164

Initial Values

 Initial Smoothed Value  : 88.263
 Initial Trend Parameter :  2.645

Seasonal Indices for the Series are

   1     87.628
   2     82.996
   3     95.362
   4     98.155
   5    101.936
   6    117.471
   7    133.389
   8    130.644
   9    107.630
  10     93.352
  11     78.956
  12     86.086
```

Final Values

```
 Final Smoothed Value   : 497.922
 Final Trend Parameter  :   8.374
 Within Series MSE      : 325.915
 Standard Error         :  18.053

 Observation   Forecast
 ------------------------
 145           443.655
 146           427.158
 147           498.784
 148           521.612
 149           550.243
 150           643.937
 151           742.367
 152           738.027
 153           617.035
 154           542.995
```



The output begins with the model and initial parameter estimates. The *Initial Smoothed Value* is a regression estimate of the level of the seasonally adjusted series immediately

before the first observation in the sample. The *Initial Trend Parameter* is the slope of the regression of observations on observation number— the increase or decrease from one observation to the next due to the overall trend. For a percentage growth model, the trend parameter is the expected percentage change from the previous to the current observation due to trend.

After the values are smoothed, SYSTAT prints the final estimates of the seasonal, location, and trend parameters, plus the within- series forecast error. You can vary the smoothing coefficients and see if they reduce the standard error.

Alternative Smoothing Coefficients

In an attempt to reduce the standard error, we alter the smoothing coefficients.

The input is:

```
CLEAR
USE AIRLINE
SERIES
   TIME 1949 12/ FORMAT = 'MMM.YYYY'
   EXPONENTIAL PASS / SMOOTH=.2  LINEAR=.2  MULT=.2 FORECAST=10
```

The output is:

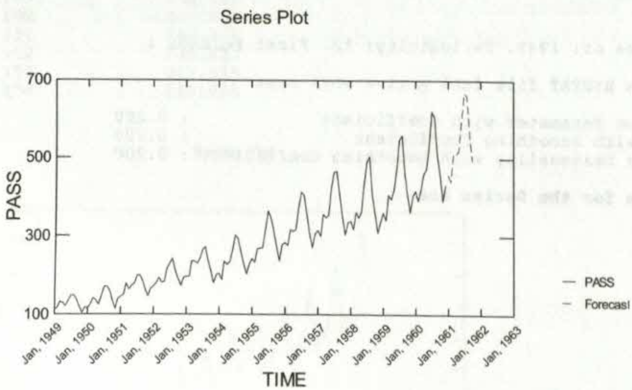
```
Series is Cleared

PASS
Series Originates at: 1949. Periodicity: 12. First Period: 1

PASS copied from SYSTAT file into active work area

  Smooth Location Parameter with Coefficient             : 0.200
  Linear Trend with Smoothing Coefficient                : 0.200
  Multiplicative Seasonality with Smoothing Coefficient  : 0.200

Seasonal Indices for the Series are

   1    91.077
   2    88.133
   3   100.825
   4    97.321
   5    98.305
   6   111.296
   7   122.636
   8   121.652
   9   105.997
  10    92.200
  11    80.397
  12    90.164

Initial Values

  Initial Smoothed Value  : 88.263
  Initial Trend Parameter :  2.645
```

```
Seasonal Indices for the Series are

  1    90.960
  2    87.249
  3    99.846
  4    98.817
  5   100.173
  6   113.664
  7   126.653
  8   124.419
  9   105.075
 10    91.727
 11    79.163
 12    88.002
```

Final Values

```
Final Smoothed Value    : 499.423
Final Trend Parameter   :   4.106
Within Series MSE       : 220.026
Standard Error          :  14.833
```

Observation	Forecast
Jan,1961	458.008
Feb,1961	442.905
Mar,1961	510.954
Apr,1961	509.745
May,1961	520.853
Jun,1961	595.666
Jul,1961	668.936
Aug,1961	662.248
Sep,1961	563.597
Oct,1961	495.771



We get a smaller within- series forecast error (220.026 versus 325.915).

In-Series Forecasts

Sometimes it is best to develop a model on a portion of a series and see how well it predicts the remainder. There are 12 years of airline data for a total of 144 monthly observations. The following commands develop the smoothing model with the first 10 years of data (120 observations) and predict the final 2 years (observations 121–144).

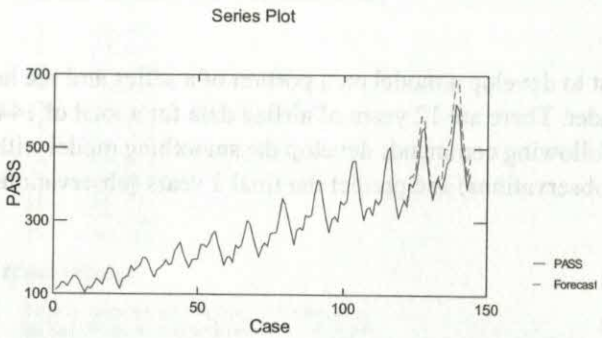
The input is:

```
CLEAR
USE AIRLINE
SERIES
  EXPONENTIAL / SMOOTH=.3 LINEAR=.4  MULT=.4 FORECAST=121 .. 144
```

Output from this procedure includes the following forecasts.

The output is:

```
Observation  Forecast
-----------------------
122          354.228
123          423.782
124          426.328
125          447.235
126          530.144
127          588.677
128          580.174
129          484.919
130          420.917
131          365.251
132          407.409
133          427.066
134          418.753
135          499.821
136          501.697
137          525.153
138          621.184
139          688.343
140          677.034
141          564.765
142          489.287
143          423.786
144          471.840

Forecast MSE    : 1904.898
Standard Error  :   43.645
```



Note that the within- series standard error is not the same as in the previous run because it is now based on only the first 120 observations. The error for the actual forecasts (18.053) is much larger than that for the in- series forecasts (17.091).

For a thorough review of issues and developments in exponential smoothing models, see Gardner (1985). For an introduction to these models, see any introductory forecasting book, such as Makridakis et al. (1997).

Example 11
ARIMA Models

The first thing to consider in modeling the *AIRLINE* passenger data is the increasing variance in the series over time. We logged the data (in the time series plot example) and found that the variance stabilized. An upward trend remained, however, so we differenced the series (in the differencing example). We now identify which ARIMA parameters we want to estimate by plotting the data in several ways. The parameters of the ARIMA model are:

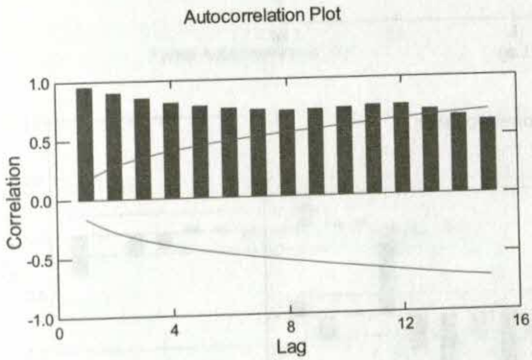
	Name	**Description**
AR	autoregressive	Each point is a weighted function of a previous point plus random error.
I	difference	Each point's value is a constant difference from a previous point's value.
MA	moving average	Each point is a weighted function of a previous point's random error plus its own random error.

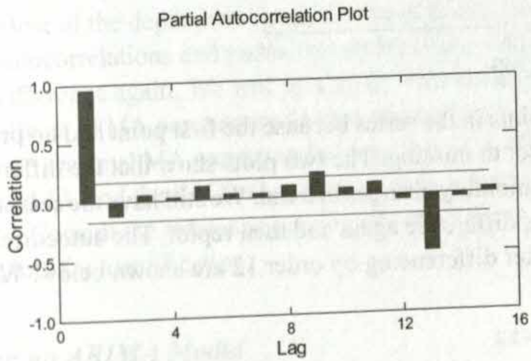
For seasonal ARIMA models, we need three additional parameters: Seasonal AR, Seasonal I, and Seasonal MA. Their definitions are the same as above, except that they

apply to points that are not adjacent in a series. The *AIRLINE* data involve seasonal parameters, for example, because dependencies extend across years as well as months.

Checking ACF and PACF Displays

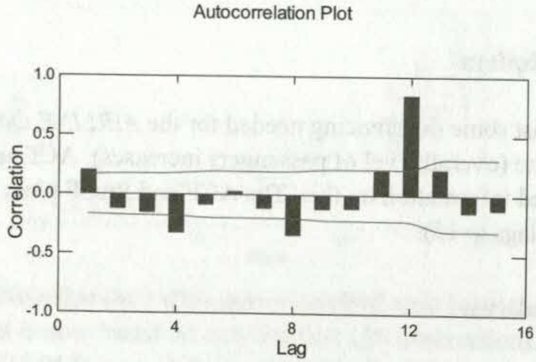
There appears to be at least some differencing needed for the *AIRLINE* data because the series drifts across time (overall level of passengers increases). ACF and PACF plots give us more detailed information on this. The ACF and PACF plots are shown below (here we limit the lags to 15).

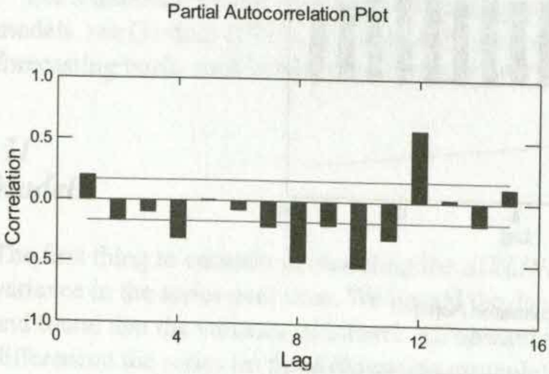




Notice that the autocorrelations are substantial and well outside two standard errors on the plot. There are two bulges in the ACF plot at lag=1 and lag=12, suggesting the nonseasonal (monthly) and seasonal (yearly) dependencies that we supposed. The

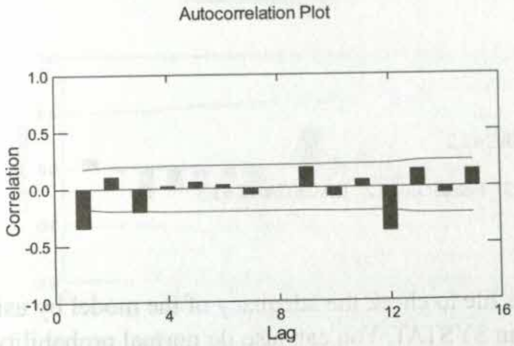
PACF plot shows the same dependencies more distinctly. Here are the autocorrelations and partial autocorrelations of the differenced series:

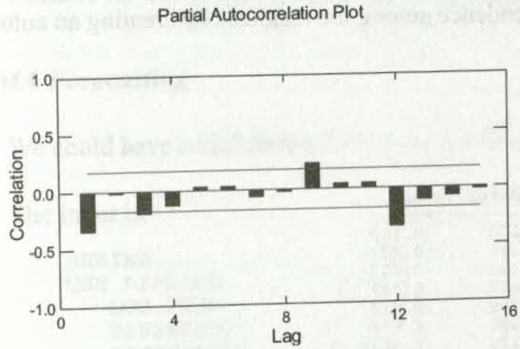




Now we have only 143 points in the series because the first point had no prior value to remove. It was therefore set to missing. The two plots show that the differencing has substantially removed the monthly changes in trend. We still have the seasonal (yearly) trend, however. Therefore, difference again and then replot. The autocorrelations and partial autocorrelations after differencing by order 12 are shown below. With commands:

```
DIFFERENCE / LAG=12
ACF / LAG=15
PACF / LAG=15
```





Most of the dependency seems to have been removed. Although there are some autocorrelations and partial autocorrelations outside two standard errors, we will not difference again. We will fit a model first because over- differencing can mask the effects of MA parameters. In fact, the pattern in this last plot suggests one regular and one seasonal MA parameter because there are ACF spikes (instead of bulges) at lags 1 and 13, and the PACF shows decay at lags 1 and 13.

Consult the references previously cited for more information on how to read these plots for identification.

Fitting an ARIMA Model

Here we fit a seasonal multiplicative ARIMA model with no autoregressive parameter, one difference parameter, one moving average parameter, no seasonal autoregressive parameter, one seasonal difference parameter, and one seasonal moving average parameter.

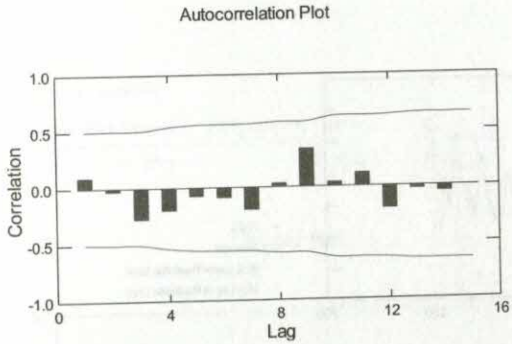
The input is:

```
SERIES
USE AIRLINE
   LOG PASS
   DIFFERENCE
   DIFFERENCE / LAG=12
   SAVE RESID
   ARIMA /Q=1 QS=1 SEASON=12 BACKCAST=13
USE RESID
   ACF / LAG=15
```

We save the residuals into a file to check the adequacy of the model by using the various facilities available in SYSTAT. You can also do normal probability plots, stem-and-leaf plots, Kolmogorov-Smirnov tests, and other statistical tests on residuals. We focus on the serial dependence among the residuals by creating an autocorrelation plot.

The output is:

```
Iteration History

 Iteration       SS     Parameter Values
-----------------------------------------
         0    0.239       0.100      0.100
         1    0.184       0.345      0.433
         2    0.176       0.449      0.633
         3    0.176       0.416      0.592
         4    0.176       0.409      0.613
         5    0.176       0.392      0.614
         6    0.176       0.396      0.613
         7    0.176       0.396      0.613
         8    0.176       0.396      0.613
         9    0.176       0.396      0.613
        10    0.176       0.396      0.613
Final Value of MSE is 0.001

Final Estimates

 Index   Type   Estimate   ASE     95% Confidence Interval
                                      Lower       Upper
-----------------------------------------------------------
   1     MA      0.396    0.093      0.212       0.579
   2     SMA     0.613    0.074      0.467       0.760

Asymptotic Correlation Matrix of Parameters

   | 1         2
---+---------------
 1 |  1.000
 2 | -0.171    1.000
```



None of the autocorrelations are significant.

ARIMA Forecasting

We could have added forecasting by specifying 10 cases to be forecast.

The input is:

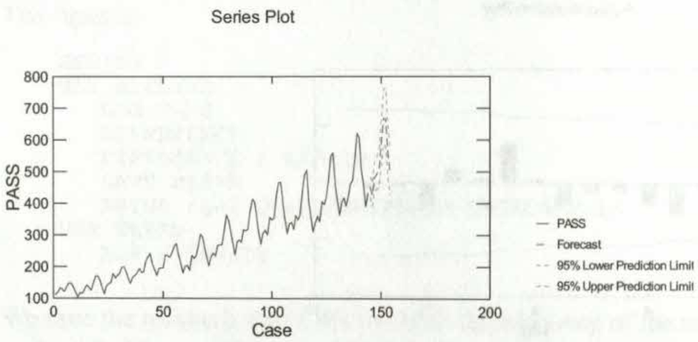
```
SERIES
USE AIRLINE
  LOG PASS
  DIFFERENCE
  DIFFERENCE / LAG=12
  SAVE RESID
  ARIMA /Q=1 QS=1 SEASON=12 BACKCAST=13 FORECAST=10
```

SYSTAT forecasts the future values of the series.

The output is:

Forecast Values

Period	Forecast	95% Prediction Interval Lower	Upper
145	450.296	418.862	484.090
146	426.557	391.976	464.189
147	482.099	438.329	530.240
148	492.247	443.297	546.602
149	508.378	453.832	569.480
150	583.440	516.623	658.898
151	668.339	587.308	760.550
152	666.091	581.137	763.465
153	558.844	484.248	644.931
154	496.754	427.648	577.028



The "forecast origin" in this case is taken as the last point in the series. From there, the model computes and prints 10 new points with their upper and lower 95% prediction intervals. SYSTAT automatically plots the forecasts.

Example 12
Mann-Kendall test

To illustrate this test, we use the value (in millions of £) of British exports during the years 1820-1850 (Hand et al., 1994). Each of the 31 yearly counts is stored in the SYSTAT file named *EXPORTS*.

The input is:

```
SERIES
USE EXPORTS
 MKTEST EXPORTS*YEAR / ALT=UPWARD SLOPE CONFI=0.95
```

The output is:

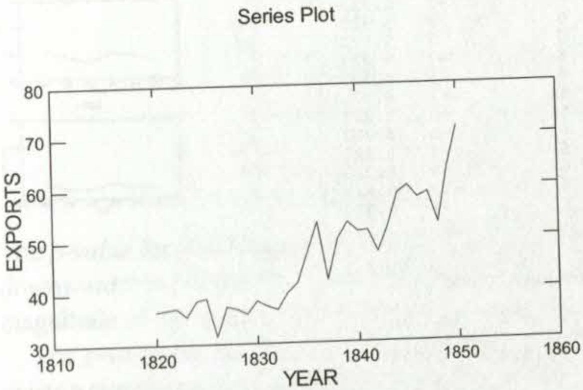
```
Series Variable                                            : EXPORTS
  Time Variable                                            : YEAR
  Number of Distinct Time Points                           : 31
  Number of Tied Group of Observations in Series Variable  : 29
```

Mann-Kendall Test

H0: No Trend vs H1: Upward Trend

Statistic	ASE	Z	p-value
345	58.819	5.848	0.000

```
Slope Estimator

 Slope Estimate   95% Lower Limit
---------------------------------
          0.975             0.813
```



The *p-value* for the Mann-Kendall test statistic indicates that there is a significant upward trend in the yearly exports. The Sen's slope estimator gives an estimate of the magnitude of the trend that was detected. If there is no trend in data, the Sen's Slope estimate should be near zero. However, the slope estimate of 0.975 indicates that there is an upward trend in the data.

Example 13
Seasonal Trend tests

To illustrate this test, we use the monthly deaths from bronchitis, emphysema and asthma in the UK during 1974-1979(Hand et al., 1994). Each of the 72 monthly counts is stored in the SYSTAT file named *LUNGDIS*.

The input is:

```
SERIES
USE LUNGDIS
  STEST DEATHS*YEAR/SEASON=MONTH$ TEST=SK MSK HT,
ALT=DOWNWARD SLOPE CONFI=0.90
```

The output is:

```
Series Variable    : DEATHS
  Time Variable      : YEAR
  Seasonal Variable : MONTH$

Mann-Kendall Statistic for Seasons

            Number of
  Season   Observations   Statistic      ASE
  ------------------------------------------
  JAN                 6          -7    5.323
  FEB                 6           5    5.323
  MAR                 6           3    5.323
  APR                 6          -7    5.323
  MAY                 6          -5    5.323
  JUN                 6          -1    5.323
  JUL                 6          -5    5.323
  AUG                 6          -5    5.323
  SEP                 6          -4    5.228
  OCT                 6          -9    5.323
  NOV                 6         -13    5.323
  DEC                 6          -3    5.323
```

Seasonal Kendall Test

```
H0: No Trend vs H1: Downward Trend

  Statistic      ASE        Z    p-value
  --------------------------------------
    -51.000   18.412   -2.824      0.002
```

Slope Estimator

```
  Slope Estimate   90% Upper Limit
  --------------------------------
         -26.000           -13.362
```

Modified Seasonal Kendall Test

```
H0: No Trend vs H1: Downward Trend

  Statistic      ASE        Z    p-value
  --------------------------------------
    -51.000   31.512   -1.650      0.049
```

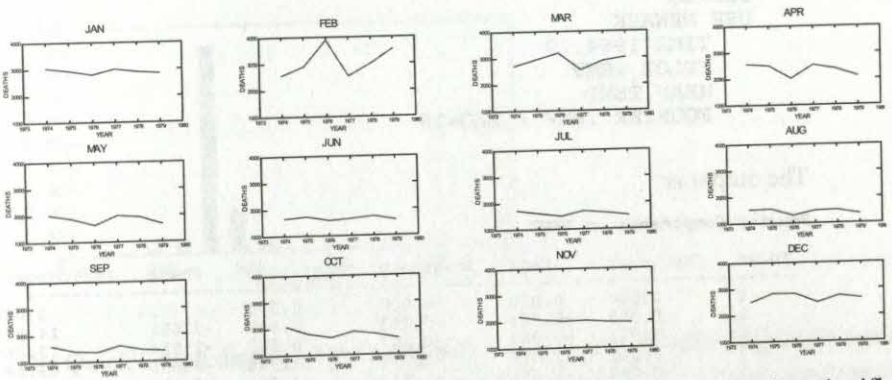
Slope Estimator

```
  Slope Estimate   90% Upper Limit
  --------------------------------
         -26.000            -7.138
```

Homogeneity Test

```
H0: Homogeneous Seasonal Trend vs H1: Non Homogeneous Seasonal Trend
H0*: No Trend vs H1*: Presence of Monotonic Trend

                 Statistic   df   p-value
  ---------------------------------------
  Homogeneity        9.396   11     0.585
  Trend              7.672    1     0.006
  Total             17.068   12     0.147
```

Seasonal Trend Plot



The *p-value* for the Seasonal Kendall test statistic indicates that there is a significant downward trend in the data and the Seasonal Kendall slope estimate of -26 gives the magnitude of the downward trend present in the data.

The *p-value* for the Modified Seasonal Kendall test statistic indicates that there exists a significant decreasing trend and the Modified Seasonal Kendall slope estimate of -26 indicates the magnitude of the downward trend in the data.

From the *p-value* corresponding to homogeneity in the Homogeneity test, there is an evidence of homogeneous seasonal trend over time. Since the homogeneity is confirmed, we go for the test of overall monotonic trend. From the *p-value* corresponding to the trend in the Homogeneity test we conclude that there is a significant monotonic trend across all the seasons.

From the seasonal trend plot we observe that the trend pattern over time is the same in most of the seasons.

Example 14
Fourier Modeling of Temperature

Let us look at a typical Fourier application. The data in the *NEWARK* file are 64 average monthly temperatures in Newark, New Jersey, beginning in January, 1964. The data are from the U.S. government, cited in Chambers et al. (1983). Notice that their fluctuations look something like a sine wave, so we might expect that they could be modeled adequately by the sum of a relatively small number of trigonometric components. We have taken exactly 64 measurements to fulfill the powers of 2 rule.

We remove the series mean before the decomposition.

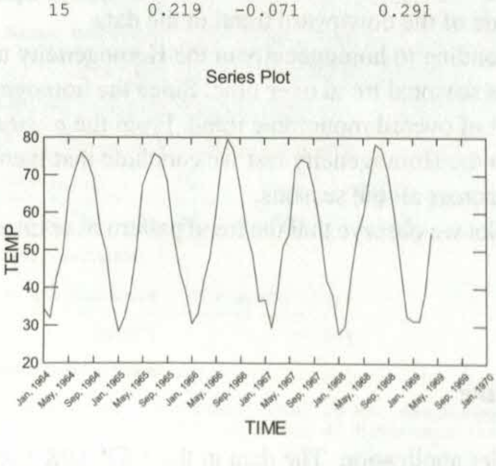
The input is:

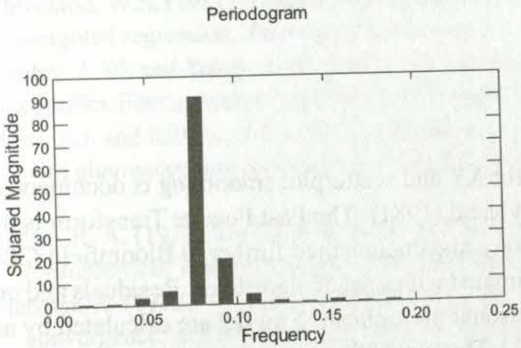
```
SERIES
USE NEWARK
   TIME 1964,12
   TPLOT TEMP
   MEAN TEMP
   FOURIER TEMP / LAG=15
```

The output is:

Fourier Components of TEMP

Index	Frequency	Real	Imaginary	Magnitude	Phase	Periodogram
1	0.000	0.000	0.000	0.000	.	0.000
2	0.016	-0.763	-0.363	0.845	-2.697	14.535
3	0.031	-0.803	-0.177	0.822	-2.924	13.760
4	0.047	-1.587	-0.779	1.768	-2.685	63.683
5	0.063	-1.658	-1.817	2.460	-2.310	123.262
6	0.078	-6.248	-7.214	9.544	-2.285	1855.632
7	0.094	2.606	3.633	4.471	0.948	407.199
8	0.109	1.040	1.786	2.067	1.044	87.038
9	0.125	0.592	0.936	1.107	1.007	24.978
10	0.141	0.438	0.588	0.733	0.930	10.954
11	0.156	-0.127	1.135	1.142	1.682	26.558
12	0.172	0.067	0.715	0.718	1.477	10.507
13	0.188	-0.255	0.785	0.825	1.885	13.860
14	0.203	0.140	0.132	0.192	0.756	0.749
15	0.219	-0.071	0.291	0.299	1.811	1.823





The Quick Graph displays a periodogram — that is, the squared magnitude against frequencies. Notice that our hunch was largely correct. There is one primary peak at a relatively low frequency. This periodogram differs from that produced in earlier versions of SYSTAT. SYSTAT now uses

```
N/pi*(squared magnitude)
```

where *N* is the number of cases in the file.

Two final points follow. First, some analysts prefer to plot the logs of these values against frequency. We could do this in the following way:

```
SQUARE TEMP
LOG TEMP
TPLOT TEMP
```

Logging, by the way, looks noisier than the plot above but can reveal significant spikes that might be hidden in the raw periodogram.

The second point involves smoothing the periodogram. Often it is best to taper the series first before computing the periodogram. This makes the spikes more pronounced in the log-periodogram plot:

```
MEAN TEMP
TAPER TEMP
FOURIER TEMP
SQUARE TEMP
LOG TEMP
TPLOT TEMP
```

Since we did not specify a value, split-cosine-bell used its default as 0.5.

Computation

Algorithms

The LOWESS algorithm for XY and scatterplot smoothing is documented in Cleveland (1979) and Cleveland (1981). The Fast Fourier Transform is due to Gentleman and Sande (1966), and documented further in Bloomfield (2000).

ARIMA models are estimated with a set of algorithms. Residuals and unconditional sums of squares for the seasonal multiplicative model are calculated by an algorithm in McLeod and Sales (1983). The sums of squares are minimized iteratively by a quasi-Newton method due to Fletcher (1972). A penalty function for inadmissible values of the parameters makes this procedure relatively robust when values are near the circumference of the unit circle. Standard errors for the parameter estimates are computed from the inverse of the numeric estimate of the Hessian matrix, following Fisher (1922). Forecasting is performed via the difference equations documented in Chapter 5 of Box et al. (1994).

References

Bennett, R.M. and Desmarais, R.N. (1975). Curve fitting of aeroelastic transient response data with exponential functions. *In Flutter Testing Techniques*.Report of a conference held at Dayton Flight Research Center, Edwards, CA, October 9-10, 1975.Washington, DC: NASA. 43-58.

Bloomfield, P. (2000). *Fourier analysis of time series: An introduction*, 2nd ed. New York: John Wiley & Sons.

Box, G. E. P., Jenkins, G. M., and Reinsel G.C. (1994). *Time series analysis: Forecasting and control*, 3rd ed, Englewood Clifs, NJ: Prentice-Hall.

Brigham, E. O. (1988). *The fast Fourier transform and its applications*. New York: Prentice-Hall.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. (1983). *Graphical methods for data analysis*. Belmont, Calif.: Wadsworth International Group.

Chatterjee, S. and Price, B. (2006). *Regression analysis by example*. 4th ed. New York: John Wiley & Sons.

Cleveland, W. S. (1979). Robust locally weight regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W.S. (1981). Lowess: A program for smoothing scatter plots by robust locally weighted regression, *American Statistician*, 35, 54.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine computation of complex Fourier series. *Mathematical Computation*, 19, 297–301.

* Dietz, E.J. and Killeen, T.J. (1981). A nonparametric multivariate test for monotone trend with pharmaceutical applications. *Journal of American Statistical Association*, 76, 169-174.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, Series A, 222, 309-368.

Fletcher, R. (1972). A class of methods for nonlinear programming, iii: Rates of convergence, in *Numerical Methods for Nonlinear Optimization* (F. A. Lootsma, ed.), New York: Academic Press, 371-382.

Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1–28.

Gentleman, W. M. and Sande, G.(1966). Fast Fourier transforms-for fun and profit. *Proc. AFIPS*, 29, 563-578.

Gilbert, R.O. (1987). *Statistical methods for environmental pollution monitoring*. New York: John Wiley & Sons.

Hand, D. J., Daly, F., Lunn A. D., McConway, K. J., and Ostrowski, E. (Editors) (1994). *A handbook of small data sets*. London: Chapman & Hall.

Hirsch, R.M., Slack, J.R., and Smith, R.A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 18(1), 107-121.

* Hirsch, R.M. and Slack, J.R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, 20, 727-732.

Makridakis, W., Wheelwright, S. C., and Hyndman, R, J. (1997). *Forecasting: Methods and applications*, 3rd ed. New York: John Wiley & Sons.

Mann, H.B. (1945). Non-parametric tests against trend, *Econometrica*, 13, 245-259.

McCleary, R. and Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. Beverly Hills: Sage Publications.

McLeod, A.I., and Sales, P.R.H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Applied Statistics*, 32, 211-223.

Nelson, C.R. (1973). *Applied time series analysis for managerial forecasting*. San Fransico: Holden-Day.

Sen, P.K. (1968). Estimates of regression coefficients based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379-1389.

Turkey, J.W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Van Belle, G. and J.P. Hughes. (1984). Nonparametric tests for trend in water quality, *Water Resources Research*, 20, 127-136.

Vandaele, W. (1983). *Applied time series and Box-Jenkins models*. New York: Academic Press.

Velleman, P.F. and Hoaglin, D.C. (1981). *Applications, basics, and computing of exploratory data analysis.* Belmont: Duxbury Press.

* (indicates additional reference)

Chapter

16

Two-Stage Least Squares

Dan Steinberg

The TSLS module is designed for the estimation of simultaneous equation systems via Two-Stage Least Squares (TSLS) and Two-Stage Instrumental Variables (White, 2000). In the first stage, the independent variables are regressed on the instrumental variables. In the second stage, the dependent variable is regressed on the predicted values of the independent variables (determined from the first stage). TSLS produces heteroskedasticity-consistent standard errors for ordinary least-squares (OLS) models and instrumental variables models and provides diagnostic tests for heteroskedasticity and nonlinearity. TSLS also computes regressions with polynomially distributed lag structures in the errors.

The polynomial distributed lag is a method of including a large number of lagged variables in a model by reducing the number of coefficients to be estimated, by requiring the coefficients to lie on a smooth polynomial in the lagged variables. A PDL (polynomial distributed lag) variable specification may be used as an independent (predictor) variable in any linear regression procedure.

TSLS also tests whether all the parameters are jointly significant or individually for each parameter by using general linear (Wald) hypothesis testing.

Statistical Background

Two-stage least squares was introduced by Theil in the early 1950's in unpublished memoranda and independently by Basmann (1957). Theil's textbook (1971) treats the topic extensively; other textbooks include Johnston (1997), Judge et al. (1988), Maddala and Lahiri (2001), and Mardia et al. (1979).

Two-Stage Least Squares Estimation

Two-Stage Least Squares (TSLS) is the most common example of an instrumental variables (*IV*) estimator. The *IV* estimator is appropriate if we want to fit the statistical model

$$y = \boldsymbol{Xb} + \varepsilon$$

when some of the regressors in $\boldsymbol{X}$ are correlated with the errors ε. This can occur if some of the $\boldsymbol{X}$'s are measured with error or when some of the $\boldsymbol{X}$'s are dependent variables in a larger system of equations.

To use the instrumental variables procedure, we must have some variables $\boldsymbol{Z}$ in our data set that are uncorrelated with the error terms ε $(E[\boldsymbol{Z}'\varepsilon] = 0)$. These variables, which are called the **instrumental variables**, can include some or all of the variables $\boldsymbol{X}$ of the model and any other variables in the data. To estimate a model, there must be at least as many instrumental variables as there are regressors.

The main objective of this Polynomial Distributed Lag method is to reduce the number of parameters (coefficients) to be estimated.

Let the general distributed lag model with a finite lag of *k* time periods be

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_k X_{t-k} + u_t$$

where u_t is the disturbance term.

In the Almon Polynomial Distributed Lag model, Almon assumes that β_i can be approximated by a suitable degree polynomial in *i*, more generally, an m^{th} degree polynomial in *i* as

$$\beta_i = a_0 + a_1 i + a_2 i^2 + \ldots + a_m i^m\text{, where } m < k.$$

Using the above polynomial of β_i, the general distributed lag model with a finite lag of *k* time periods can be re-written as:

$$Y_t = \alpha + a_0 Z_{0,t} + a_1 Z_{1,t} + a_2 Z_{2,t} + \ldots + a_k Z_{k,t} + u_t$$

where

$$Z_{0,t} = \sum_{i=0}^{k} X_{t-i} \quad \text{and} \quad Z_{j,t} = \sum_{i=0}^{k} i^j X_{t-i} \quad \text{for } j = 1, 2, 3, \ldots k$$

Y is regressed on the constructed variables $Z_{0,t}$, $Z_{j,t}$, etc., and the parameters a_i's can be estimated by the usual OLS procedure, provided that $Z_{i,t}$ and u_t are uncorrelated. If $Z_{i,t}$ and u_t are correlated, then the parameters a_i's can be estimated by using the Two-Stage Least Squares method by identifying appropriate instrumental variable(s).

Heteroskedasticity

The problem of heteroskedasticity is discussed in Theil (1971) and extensively in Judge et al. (1988), which includes numerous references. The approach to heteroskedasticity taken in this module, which is to produce correct standard errors for the OLS case, was introduced by Eicker (1963, 1967) and Hinkley (1977). It was rediscovered independently by White (1980), who also extended its application to the TSLS context (White, 1982a,1982b). A technical account of the theory underlying all of the methods used in this module appears in White (2000). The basic statistical model of regression analysis can be written as:

$$y = \boldsymbol{X}\boldsymbol{b} + \varepsilon$$

where y is the dependent variable, $\boldsymbol{X}$ is a vector of independent variables, $\boldsymbol{b}$ is a vector of unknown regression coefficients, and ε is an unobservable random variable. If the regressors are uncorrelated with the random error $(\mathrm{E}\,[\boldsymbol{X}'\varepsilon] = 0)$, ordinary least-squares (OLS) will generally produce consistent and asymptotically normal estimators. Further, if the errors have constant variance for all of the observations in the data set, the usual t statistics are correct and hypothesis testing can be conducted on the basis of the variance-covariance matrix of the coefficient estimates.These are the assumptions underlying the estimation and hypothesis testing of GLM and other major regression packages. If either of these assumptions is false, features of TSLS can be used to obtain valid hypothesis tests and consistent parameter estimates.

We estimate heteroskedasticity-consistent standard errors because they are correct asymptotically under a broad set of assumptions. If the random errors in a regression model exhibit heteroskedasticity, the conventional standard errors and the covariance

matrix are usually inconsistent. The t statistics are erroneous, and any hypothesis tests that employ the covariance matrix estimate will also be incorrect (have the wrong size). The heteroskedasticity-consistent standard errors, by contrast, are correct, whether or not heteroskedasticity is present.

There is no way to tell whether the robust standard errors will be larger or smaller than the OLS results, but they may differ substantially. The classical approach to heteroskedasticity is to postulate an exact functional form for the second moments of the errors. Some analysts assume, for example, that the variance of the error for each observation is proportional to the square of one of the independent variables. (See Judge et al. for further details). The model is estimated by generalized (or weighted) least-squares (GLS) with weights obtained from least-square residuals. Of course, this approach requires that the assumptions of the analyst be correct. If these assumptions are incorrect, the standard errors resulting from GLS will also be incorrect.

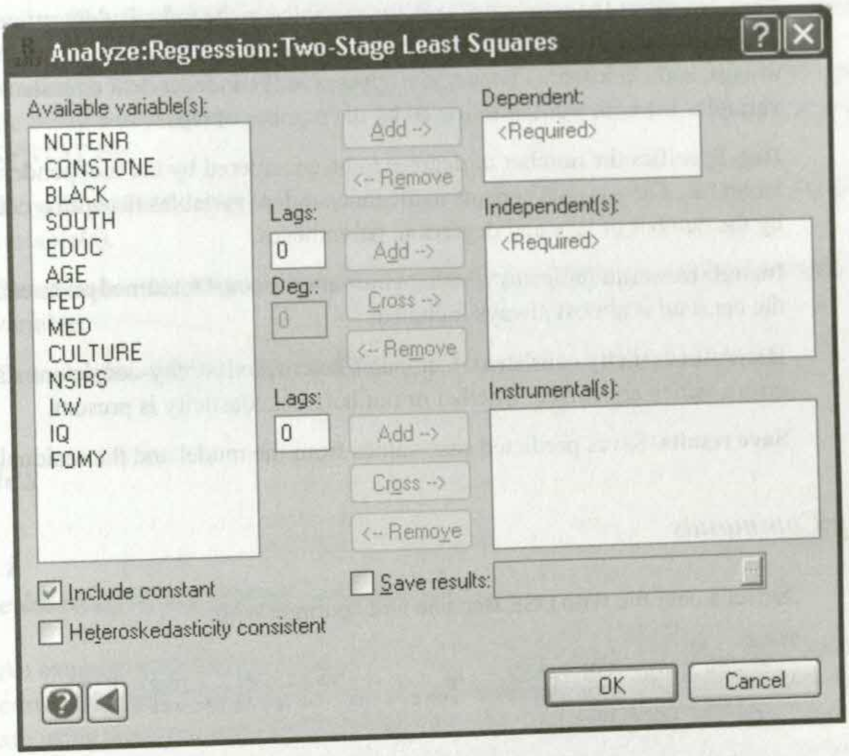
The heteroskedasticity-consistent standard errors computed in TSLS are not based on any attempt to correct for, or otherwise model, the heteroskedasticity. Instead, essentially nonparametric estimates of the OLS standard errors are computed. We still get OLS coefficients, but the variances of the coefficients are revised. White (1980) showed that this is possible for virtually any type of heteroskedasticity.

Two-Stage Least Squares in SYSTAT

Two-Stage Least Squares Regression Dialog Box

To open the Two-Stage Least Squares dialog box, from the menus choose:

Analyze
  Regression
    Two-Stage Least Squares...



SYSTAT computes Two-Stage Least Squares by first specifying a model and then estimating it.

Dependent. The variable you want to examine. The dependent variable should be continuous and numeric.

Independent(s). Selects one or more continuous or categorical variables (grouping variables).To add an interaction to your model, click the Cross button. For example, to add the term *sex*education*, add *sex* to the Independent(s) list and then add *education* by clicking the Cross button.

Instrumental(s). Selects the instrumental variable(s) that you want to estimate. Instrumentals may be continuous or categorical. To add an interaction to your model, click the Cross button. For example, to add the term *sex*education*, add *sex* to the Instrumental(s) list and then add *education* by clicking the Cross button. The number of instrumental variables must be equal or exceed the number of independent variables.

Lags. Specifies the number of lags for variables in the independent or instrumental variable target list. Highlight the variable in the general variable list, enter the number of lags, and click Add. The variable appears in the independent or instrumental variables list with a colon followed by the number of lags.

Deg. Specifies the number of degrees to be considered by the independent variable target list. The variable appears in the independent variables list with a colon followed by the number of lags and degrees in parentheses.

Include constant. Indicates whether you want the constant turned on or off. In practice, the constant is almost always included.

Heteroskedasticity consistent. Computes heteroskedasticity-consistent standard errors, which are correct whether or not heteroskedasticity is present.

Save results. Saves predicted case values from the model and the residual to a *file.*

Using Commands

Select a data file with USE *filename* and continue with:

```
TSLS
  MODEL yvar = CONSTANT + var + ... + var1*var2 + pvar:k(m) +...|,
               CONSTANT + ivar + ... + ivar1*ivar2 + ...
  ESTIMATE / HC
```

The term *pvar:k(m)* is used to include a PDL term of lag *k* and degree **m.**

The CONSTANT term is almost always included. The second block of variables is the set of predictors (regressors). Any predictor can be declared categorical with the CATEGORY statement. This will encode them with dummy variables. After a vertical bar, list the instrumental variables.

Usage Considerations

Types of data. TSLS uses rectangular data.

Print options. PLENGTH NONE gives the table of *t*-ratios. The output is standard for the other PLENGTH options.

Quick Graphs. No Quick Graphs are produced by TSLS.

Saving files. Predicted values and residuals can be saved to a SYSTAT system file with the SAVE command. The SAVE command must be issued before the ESTIMATE command. The SAVE command works across the BY command. If several BY groups are being analyzed, the predicted values, etc., are saved for each BY group in a single SYSTAT file.

BY groups. TSLS analyzes data by groups. Your file need not be sorted on the BY variable(s).

Case frequencies. FREQ <variable> increases the number of cases by the FREQ variable.

Case weights. WEIGHT is not available in TSLS.

Examples

Example 1
Heteroskedasticity-Consistent Standard Errors

An example will illustrate how to use TSLS to diagnose a regression model and obtain correct answers if the classical regression assumptions are violated. The data that we are using were extracted from the National Longitudinal Survey of Young Men, 1979. Information for 38 men is available on natural log of wage (*LW*), highest completed grade (*EDUC*), mother's education (*MED*), father's education (*FED*), race (*BLACK*=1), *AGE*, and several other variables. We want to estimate a model relating wage to education, race, and age for a simple linear model with heteroskedasticity-consistent standard errors.

The input is:

```
TSLS
   USE NLS
   MODEL LW=CONSTANT+EDUC+BLACK+AGE
   ESTIMATE / HC
```

The output is:

```
Input Records                : 200
Records Kept for Analysis : 200

Ordinary Least Squares (OLS) Results

N                                        : 200.000000
Mean of Dependent Variable               :   6.080000
R-squared                                :   0.140278
Adjusted R-squared                       :   0.127119
Uncentered R-squared (R0-squared)        :   0.994372
```

Regression Coefficients

Parameter	Estimate	Standard Error	t	p-value
CONSTANT	4.483	0.327	13.726	0.000
EDUC	0.023	0.013	1.715	0.088
BLACK	-0.207	0.135	-1.530	0.128
AGE	0.050	0.011	4.605	0.000

```
F-ratio : 10.660242
df      :    (3,196)
p-value : 0.000002
```

Standard Error of Regression	0.462279
Regression Sum of Squares	6.834354
Residual Sum of Squares	41.885646

Covariance Matrix of Regression Coefficients

	1	2	3	4
1	0.107			
2	-0.002	0.000		
3	-0.009	0.000	0.018	
4	-0.003	0.000	0.000	0.000

Heteroskedastic Consistent Results

Parameter	Estimate	Standard Error	t	p-value
CONSTANT	4.483	0.335	13.386	0.000
EDUC	0.023	0.014	1.633	0.104
BLACK	-0.207	0.163	-1.266	0.207
AGE	0.050	0.012	4.254	0.000

Specification Test	Statistic	df	p-value
Durbin-Watson	1.836		
White Specification	7.907	8.000	0.443
Nonlinearity	7.710	5.000	0.173

Heteroskedastic Consistent Covariance Matrix of Regression Coefficients

	1	2	3	4
1	0.112			
2	-0.002	0.000		
3	-0.015	0.000	0.027	
4	-0.003	0.000	0.000	0.000

The initial TSLS output reports conventional regression output, which could also have been obtained in GLM.The standard errors reported are obtained from the diagonal of the matrix $s^2(X'X)^{-1}$, where s^2 is the sum of squared residuals divided by the degrees

of freedom; the t statistics are the classical ones as well. The one new statistic is the uncentered R^2, reported as *R0-squared*(R_o^2). This statistic has no bearing on the goodness of fit of the regression and should be routinely ignored. It is reported only as a computational convenience for those who want to use the Lagrange multiplier tests discussed by Engle (1984).

In addition to producing what White (1980) called heteroskedasticity-consistent standard errors, the TSLS module calculates three diagnostic statistics for the linear model. The first is the usual Durbin-Watson statistic. This is the same test for autocorrelation that appears in the GLM module.

The second is the White (1980) specification test, which explicitly checks the residuals for heteroskedasticity. Under the null hypothesis of homoscedasticity, this statistic has an asymptotic chi-square distribution. If this statistic is large, we have evidence of heteroskedasticity. In the example, this statistic is 7.907 with eight degrees of freedom, indicating that we cannot reject the null hypothesis of homoscedasticity. The White test is actually a general test of misspecification (White, 1982) and is sensitive to various departures of the model and data from standard assumptions. A significant statistic is an evidence for something wrong with the model, but it does not identify the source of the problem. For example, it may be heteroskedasticity, but it could also be left-out variables or nonlinearity.

The third is the nonlinearity test, which checks for neglected nonlinearities in the regression function. It is simply a test of the joint hypothesis that all possible interactions, including squared regressors, have zero coefficients in a full model. A Lagrange multiplier test also has an approximate chi-square distribution under the null hypothesis of correct specification. Again, large values for this statistic are evidence for neglected interactions. In our example, there is no evidence of nonlinearity on the basis of this broad test.

The latter two tests involve supplementary regressions with a possibly large number of additional independent variables. The White test is computed by regressing the squared OLS residuals on all the squares and cross-products of the X's, and the nonlinearity test regresses residuals on these same cross-products.

Example 2
Two-Stage Least Squares

In this example, we use the extended MODEL statement to construct a TSLS model.

The input is:

```
TSLS
  USE NLS
  MODEL LW=CONSTANT+EDUC+BLACK+AGE | CONSTANT+MED+FED+BLACK+AGE
  ESTIMATE
```

The first part of the MODEL statement is identical to what we have seen in GLM. It has a dependent variable and a list of independent variables, characterizing a structural equation. This is the theoretical model we want to estimate. The vertical bar (|) in the middle of the statement signifies the end of the structural equation, followed by the list of instrumental variables. In this example, our structural equation relates the logarithm of the wage, *LW*, to *EDUC, BLACK, AGE*, and a *CONSTANT*. The list of instrumental variables follows, and it consists of *CONSTANT, MED, FED, BLACK,* and *AGE*. At this point, all we need to know is that there are at least as many instrumental variables to the right of the sharp sign as there are regressors to the left. Satisfying this condition means that TSLS will attempt to fit the model.

What exactly does our model mean? In our example, *CONSTANT*, *BLACK*, and *AGE* appear both in the structural equation and in the list of instrumental variables. By using them in this way, the analyst expresses confidence that these are conventional independent variables, uncorrelated with the error term ε. These exogenous variables can be said to be instruments for themselves. The variable *EDUC*, however, appears only in the structural equation. This is a signal that the analyst wants to consider *EDUC* as an endogenous variable that might be correlated with the error term. On the right hand side of the vertical bar, *MED* and *FED* appear as instrumental variables only. They can thus be said to be instruments for *EDUC*.

The total number of instruments is five, which is one greater than the number of regressors. Notice that if the lists before and after the vertical bar are identical, the procedure reduces mathematically to OLS; TSLS, however, will do a lot of extra work to discover this.

Some analysts prefer to think of Two-Stage Least Squares as involving a literal pair of estimated regressions. From this point of view, we have, for example, the following two equations:

```
MODEL LW = CONSTANT + EDUC + BLACK + AGE
```

```
MODEL EDUC = CONSTANT + FED + MED
```

The first equation is the structural equation for *LW*, and the second is a possible structural equation for *EDUC*, relating *EDUC* to the education levels of parents. Because *EDUC* is itself seen to be a dependent variable in the larger set of equations, it cannot properly appear as a regressor in a standard regression. The two-stage technique involves estimating the equation for *EDUC* first, forming predicted values for *EDUC*, say *EDUCHAT*, and then estimating the model,

```
MODEL LW = CONSTANT + EDUCHAT + BLACK + AGE
```

instead. However, to estimate a TSLS model in a literal pair of regressions, the equation for *EDUC* would have to expand to include all of the exogenous independent variables appearing in the equation for *LW*. That is, the correct first-stage regression would actually be,

```
MODEL EDUC = CONSTANT + FED + MED + BLACK + AGE
```

although we thought the shorter model was the "true" model. Also, the standard errors obtained from a literal two-stage estimation are not correct, as they must be calculated from actual and not predicted values of the independent variables. Fortunately, these details are taken care of by TSLS. Just make sure to partition your variables into exogenous and endogenous groups and list all the exogenous variables to the right of the vertical bar.

The output is:

```
Input Records                : 200
Records Kept for Analysis    : 200

Instrumental Variables, OLS Results (TSLS)

N                                         : 200.000000
Mean of Dependent Variable                :   6.080000

Regression Coefficients

Parameter | Estimate   Standard Error       t    p-value
----------+---------------------------------------------
CONSTANT  |  4.695          0.491        9.564    0.000
EDUC      |  0.006          0.033        0.177    0.860
BLACK     | -0.223          0.138       -1.610    0.109
AGE       |  0.051          0.011        4.614    0.000

Standard Error of Regression |  0.464223
```

```
Residual Sum of Squares     | 42.238554

Covariance Matrix of Regression Coefficients
  |     1        2        3        4
--+-------------------------------
1 |  0.241
2 | -0.013   0.001
3 | -0.020   0.001    0.019
4 | -0.002   0.000    0.000    0.000
```

Example 3
Two-Stage Instrumental Variables

As in the case of the OLS estimator, the standard errors calculated for the TSLS estimator will be incorrect in the presence of heteroskedasticity. Although we could calculate heteroskedasticity-consistent standard errors, it turns out that sometimes we can do even better. If the number of instrumental variables is strictly greater than the number of regressors in the presence of heteroskedasticity, the two-stage instrumental variables (*TSIV*) estimator is more efficient than TSLS. This means that in *TSIV*, the coefficient estimates as well as the standard errors may differ somewhat from TSLS. Observe, though, that large differences between *TSIV* and TSLS coefficients may indicate model misspecification (for example, the variables assumed to be exogenous are not truly exogenous). As in the case of heteroskedasticity-consistent OLS, computation of the *TSIV* estimator does not require knowledge of the form of heteroskedasticity. See White (1982, 2000) for more on the *TSIV* estimator.

The following sequence of statements tells TSLS to estimate the model by *TSIV* as well as by TSLS. The only difference from TSLS is that the HC option is requested.

The input is:

```
TSLS
   USE NLS
   MODEL LW=CONSTANT+EDUC+BLACK+AGE | CONSTANT+MED+FED+BLACK+AGE
   ESTIMATE / HC
```

If the number of instruments is the same as the number of regressors, the *TSIV* and TSLS coefficient estimators are identical. In this case, only the standard errors printed under the *TSIV* results will differ from TSLS. These are the heteroskedasticity-consistent standard errors for the usual *IV* estimator.

The output is:

```
Input Records             : 200
Records Kept for Analysis : 200

Instrumental Variables, OLS Results (TSLS)

N                                              : 200.000000
Mean of Dependent Variable                     :   6.080000
```

Regression Coefficients

```
Parameter | Estimate  Standard Error        t  p-value
----------+------------------------------------------
CONSTANT  |    4.695           0.491    9.564    0.000
EDUC      |    0.006           0.033    0.177    0.860
BLACK     |   -0.223           0.138   -1.610    0.109
AGE       |    0.051           0.011    4.614    0.000

Standard Error of Regression |  0.464223
Residual Sum of Squares      | 42.238554
```

Covariance Matrix of Regression Coefficients

```
  |      1        2        3        4
--+----------------------------------
1 |  0.241
2 | -0.013    0.001
3 | -0.020    0.001    0.019
4 | -0.002    0.000    0.000    0.000
```

```
Instrumental Variables, Heteroscedastic Consistent Results (2SIV)

Parameter   Estimate  Standard Error        t  p-value
CONSTANT       4.612           0.514    8.976    0.000
EDUC           0.020           0.032    0.610    0.542
BLACK         -0.156           0.169   -0.924    0.357
AGE            0.046           0.012    3.995    0.000
```

Covariance Matrix of Regression Coefficients

```
  |      1        2        3        4
--+----------------------------------
1 |  0.264
2 | -0.014    0.001
3 | -0.031    0.001    0.029
4 | -0.003    0.000    0.000    0.000
```

Example 4
TSLS without lag and with hypothesis testing

The data set used here is *PDLEX2*. This data set relates to the Sales and Inventory of a product for the United States for the period 1954-1999. Let us consider a model with Sales as the independent variable and Inventory as the dependent variable.

The input is:

```
TSLS
  USE PDLEX2
  MODEL INVENTORY = CONSTANT + SALES
  ESTIMATE
  HYPOTHESIS
  TEST
```

The output is:

```
Input Records                : 46
Records Kept for Analysis : 46

Ordinary Least Squares (OLS) Results
N                                        :      46.000000
Mean of Dependent Variable               :  218913.087000
R-squared                                :       0.976228
Adjusted R-squared                       :       0.975688
Uncentered R-squared (R0-squared)        :       0.992534

 Regression Coefficients

 Parameter | Estimate   Standard Error      t    p-value
-----------+--------------------------------------------
 CONSTANT  | 23969.174        5734.739   4.180     0.000
 SALES     |     1.429           0.034  42.508     0.000

F-ratio : 1806.923051
df      :      (1,44)
p-value :    0.000000

 Standard Error of Regression |  23352.574541
 Regression Sum of Squares    | 9.853924E+011
 Residual Sum of Squares      | 2.399508E+010

 Covariance Matrix of Regression Coefficients

   |         1        2
---+---------------------
 1 | 32887234.715
 2 |     -154.200   0.001

Entering Hypothesis Procedure
No Constraints Specified. Using Default System that
all Parameters are Jointly Zero.

 Linear Restriction System

 EQN |              Parameter
     | 1        2        RHS          Q
-----+--------------------------------------
   1 | 1.000    0.000    0.000    23969.174
   2 | 0.000    1.000    0.000        1.429

General Linear Wald Test Results

Chi-square Statistic : 5849.253
df                   :        2
p-value              :    0.000
```

Example 5
PDL without Instrumental Variables

The data set used here is *PDLEX1*. This data set relates to the Sales and Inventory of a product in 20 days. Let us consider a model with three lags and degree 2.

The input is:

```
TSLS
  USE PDLEX1
  MODEL INVENTORY = CONSTANT + SALES:3(2)
  ESTIMATE
```

The output is:

```
Input Records                                    :     20
Records Kept for Analysis                        :     17
Records Deleted for Missing Incomplete Data :     3

Ordinary Least Squares (OLS) Results

N                                                :    17.000000
Mean of Dependent Variable                       : 81869.000000
R-squared                                        :     0.996797
Adjusted R-squared                               :     0.996058

 Regression Coefficients

 Parameter | Estimate   Standard Error       t    p-value
-----------+------------------------------------------------
 CONSTANT  | -7140.754        1992.988   -3.583     0.003
 SALES     |     0.661           0.165    3.996     0.002
 SALES<1>  |     1.131           0.180    6.284     0.000
 SALES<2>  |     0.737           0.164    4.485     0.001
 SALES<3>  |    -0.522           0.235   -2.223     0.045

F-ratio : 1348.639079
df      :    (3,13)
p-value :    0.000000

 Standard Error of Regression |    1757.457568
 Regression Sum of Squares    | 1.249645E+010
 Residual Sum of Squares      | 4.015254E+007

 Covariance Matrix of Regression Coefficients

   |           1         2         3         4         5
---+------------------------------------------------------
 1 | 3971999.635
 2 |     139.029     0.027
 3 |     -23.656    -0.021     0.032
 4 |    -110.861    -0.026     0.027     0.027
 5 |    -122.584     0.014    -0.038    -0.025     0.055
```

Example 6
PDL with Instrumental Variables

The data set used here is *PDLEX3*. This data set relates to an income-money supply model. Let us consider this model with two lags and degree 1.

The input is:

```
TSLS
   USE PDLEX3
   MODEL M2 = CONSTANT + GDP:2(1) |CONSTANT + GPDI + FEDEXP + TB6
   ESTIMATE
```

The output is:

```
Input Records                                        :    30
Records Kept for Analysis                            :    28
Records Deleted for Missing Incomplete Data :         2

Instrumental Variables, OLS Results (TSLS)

N                                               :    28.000000
Mean of Dependent Variable                      : 2511.514286
```

Regression Coefficients

Parameter	Estimate	Standard Error	t	p-value
CONSTANT	-2326.592	168.344	-13.820	0.000
GDP	-0.313	0.300	-1.044	0.307
GDP<1>	0.292	0.013	23.343	0.000
GDP<2>	0.898	0.317	2.833	0.009

```
Standard Error of Regression |      200.933603
Residual Sum of Squares      |  1009357.817000
```

Covariance Matrix of Regression Coefficients

	1	2	3	4
1	28339.849			
2	8.768	0.090		
3	-1.775	-0.002	0.000	
4	-12.318	-0.095	0.003	0.100

Computation

Algorithms

TSLS computes least-squares estimates via standard high-precision algorithms. Specific details are given in the references.

Missing Data

Cases with missing data on any variable in the model are deleted before estimation.

References

Basmann, R.L. (1957). A generalized classical method of linear estimation of Coefficients in a structural equation, *Econometrica*, 25, 77-83.

*Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica,* 47, 1287–1294.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of mathematical statistics,* 34, 447–456.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* Vol. 1. Berkeley: University of California Press.

Engle, R. F. (1984). Wald, likelihood ratio and Lagrange multiplier tests in econometrics. In Griliches, Z. and Intrilligator, M. D. (eds.), *Handbook of econometrics,* Vol. II. New York: Elsevier.

* Gujarati, D.N. (2003). *Basic Econometrics*, 4th ed. New York: McGraw-Hill.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics,* 19: 285–292.

Johnston, J. (1997). *Econometric methods,* 4th ed. New York: McGraw-Hill.

Judge, G. G., Hill, R. C. Griffiths, W. E., Lutkepohl, H., and Lee, T. C. (1988). *Introduction to the theory and practice of econometrics,* 2nd ed. New York: John Wiley & Sons.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics,* 29, 305–325.

Maddala, G. S. and Lahiri, K. (2006). *Introduction to econometrics*, 3rd ed. Chichester: John Wiley & Sons.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press.

Theil, H. (1971). *Principles of econometrics*. New York: John Wiley & Sons.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica,* 48, 817–838.

White, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica,* 50, 1–25.

White, H. (1982b). Instrumental variables estimation with independent observations. *Econometrica,* 50, 483–500.

White, H. (2000). *Asymptotic theory for econometricians*. New York: Academic Press.

(* indicates additional references)

Acronym & Abbreviation Expansions

A
ABS - absolute value
ACF - autocorrelation function
ACOLOR - color axes
ACS - arccosine
ACT - actuarial life table
AD test - Anderson Darling test
ADDTREE - additive trees
ADFG - asymptotically distribution free estimate biased, Gramian
ADFU - asymptotically distribution free estimate unbiased
ADJSEASON - seasonal adjustment
AHMAX - maximum extent
AHMIN - minimum extent
AIC - Akaike information criterion
AID - automatic interaction detection
ALT - alternative
ANCOVA - analysis of covariance
ANG1 - deviation of angles from north in a clockwise direction
ANG2 - deviation of angles from horizontal (for 3D models)
ANG3 - tilt angle
ANOVA - analysis of variance
ANOVAHYPO - hypothesis tests in analysis of variance
AR - autoregressive
ARIMA - autoregressive integrated moving average
ARL - average run length
ARMA - autoregressive moving average
ARS - adaptive rejection sampling
ASCII - American Standard Code for Information Interchange
ASE - asymptotic standard error
ASN - arcsine
ATH - arc hyperbolic tangent
ATN - arctangent
AVERT - vertical extent
AVG - average

B
BC - Bray-Curtis similarity measure
BCa - Bias Corrected and accelerated
BCF - Beta cumulative function
BDF - Beta density function
BETACORR - beta correction
BIC - Bayesian information criterion
BIF - Beta inverse function
BMP - Windows bitmap
BOF - beginning-of-file
BOG - beginning-of-BY group
BONF - Bonferroni
BOOT - bootstrap
BRN - Beta random number

C
CART - classification and regression trees
CBSTAT - column basic statistics
CCF - Cauchy cumulative function
CCF - cross-correlation function
CDF - Cauchy density function
cdf/CF - cumulative distribution function
CDFUNC - coefficients for canonical variables

CFUNC - coefficients for the classification functions
CGM - Computer graphics metafile: binary or clear text
CHAZ - cumulative hazard
CHISQ - Chi-square distribution
CHOL - Cholesky decomposition
CI - confidence interval
CIF - Cauchy inverse function
CIM - confidence interval of mean
CLASS - classification
CLSTEM - stem and leaf plot for column
CMeans - canonical scores of group means
CMULTIVAR - multiple string variables
COEF - coefficients
COL/col - column
COLPCT - Column percentages
CONFIG - configuration
CONT - Contingency coefficient
CONV - convergence
CORAN - correspondence analysis
CORR - correlations
CORR1 - single correlation coefficient
CORR2 - equality of two correlations
COV - covariance
Cp - process capability index
CPL - process capability based on lower specification limit
CPU - process capability based on upper specification limit
Cpk-Process capability index for off-centered process
CR - confidence region
CRA - cost of response above UTL
CRB - cost of response below LTL
CRN - Cauchy random number
CSCORE - canonical scores
CSIZE - size of characters
CSQ - Chi-square
CSTATISTICS - column statistics
CSV - comma separated values
CUSUM - cumulative sum
CUSUM HI - Upper cumulative sum
CUSUM LO - Lower cumulative sum
CV - coefficient of variation
CVI - cross validation index

D

DBF - Dbase files
DC - deciles of risk
DECF - Double exponential cumulative function
DEDF - Double exponential density function
DEIF - Double exponential inverse function
DENFUN - density function
dep. - dependent
DERN - Double exponential random number
DET - determinant
DEVI - deviates (observed values - expected values)
DEXP - Double exponential distribution
df - degrees of freedom
DF - distribution function
DHAT - estimated distance
DIF - data interchange format
DIM - dimension
DISCRIM - discriminant analysis
DIST - distance
DIT - dot histogram
DOE - design of experiments
DOS - disc operating system
DPMO - defects per million opportunities
DPU - defects per unit
DTA - Stata files
DUCF - Discrete uniform cumulative function
DUDF - Discrete uniform density function
DUIF - Discrete uniform inverse function
DUNIFORM - Discrete uniform
DURN - Discrete uniform random number
DWLS - distance weighted least-squares

E

ECF - Exponential cumulative function

EDF - Exponential density function
EEXP - extreme value exponential
EIF - Exponential inverse function
EIGEN - eigenvalues
ELAMBDA - exp(lambda)
EM - expectation-maximization
EMF - Windows enhanced metafile
ENCF - Logit normal cumulative function
ENDF - Logit normal density function
ENIF - Logit normal inverse function
ENORMAL - Logit normal
ENRN - Logit normal random number
EOF - end-of-file
EOG - end-of-BY group
EPS - Encapsulated postscript
ERN - Exponential random number
ES - exhaustive search
ESS - error sum of squares
EW - extreme value Weibull
EWMA - exponentially weighted moving average
EXP/exp - exponential/ expected

F
FAR - false-alarm rates
FCF - F cumulative function
FCOLOR - color foreground
FDF - F density function
FIF - F inverse function
FINV - inverse of the F cumulative
FITC - fitting distribution: continuous
FITD - fitting distribution: discrete
FITDIST - fitting distributions
Flexibeta - flexible beta
FPLOT - function plots
FRN - F random number
FTD - folded trellis detector
FTDEV - Freeman-Tukey deviate
FULLCOND - full conditional
FUN - function

G
GCF - Gamma cumulative function
GCOR - groupwise correlation matrix
GCOV - groupwise covariance matrix
GCV - generalized cross validation
GDF - Gamma density function
GECF - Geometric cumulative function
GEDF - Geometric density function
GEIF - Geometric inverse function
GEN - general Toeplitz structure
GERN - Geometric random number
GG - Greenhouse Geisser
GIF - Gamma inverse function
GIF - Graphics Interchange Format
GLM - generalized linear models
GLMHYPO - hypothesis tests in general linear model
GLMPOST - post hoc estimate for repeated measures in general linear model
GLS - generalized least-squares
GMA - geometric moving average
GN - Gauss-Newton method
GOCF - Gompertz cumulative function
GODF - Gompertz density function
GOIF - Gompertz inverse function
GORN - Gompertz random number
GRN - Gamma random number
GUCF - Gumbell cumulative function
GUDF - Gumbell density function
GUIF - Gumbell inverse function
GURN - Gumbell random number

H
H & L - Hosmer and Lemeshow
HC - heteroscedasticity-consistent
HCF - Hypergeometric cumulative function
HDF - Hypergeometric density function
HF- Huynh-Feldt
HGEOMETRIC - hypergeometric
HIF - Hypergeometric inverse function
HIST - histogram
HKB - Hoerl, Kennard, and Baldwin

H-L trace - Holding-Lawley trace
HR - hit-rates
HRN - Hypergeometric random number
HSD - honestly significant differences
HTERM - terms tested hierarchically
HTML - hyper text markup language
HYMH - hybrid Metropolis-Hastings

I
IF - Inverse cumulative distribution function
IGAUSSIAN - inverse Gaussian
IGCF - Inverse Gaussian cumulative function
IGDF - Inverse Gaussian density function
IGIF - Inverse Gaussian inverse function
IGRN - Inverse Gaussian random number
IIDMC - independently and identically distributed Monte Carlo
IMPSAMPI - importance sampling integration
IMPSAMPR - importance sampling ratio
I-MR - individual and moving range
Ind/indep - independent
IndMH - Independent Metropolis-Hastings
INDSCAL - individual differences scaling
INITSAMP - initial sample
INTEG FUN - integrated function
IPA - iterated principal axis
ITER - iterations

J
JACK - jackknife
JCLASS - jackknifed classification
JMP - JMP v3.2 data files
JPEG/JPG - joint photographic experts group

K
K-M - Kaplan-Meier
KNBD - kth nearest neighborhood
KRON - Kronecker product
K-S test - Kolmogorov-Smirnov test
KS1 - one sample Kolmogorov-Smirnov tests
KS2 - two sample Kolmogorov-Smirnov tests

L
LAD - least absolute deviations
LB - larger the better
LCF - Logistic cumulative function
LCHAZ - log cumulative hazard
LCL - lower control limit
LCONV - log-likelihood convergence criteria
LDF - Logistic density function
LGM - log gamma
LGST - logistic
LIF - Logistic inverse function
L-L/LL - log likelihood
LMS- least median of squares
LMSREG - least median of squares regression
LNCF - Lognormal cumulative function
LNDF - Lognormal density function
LNIF - Lognormal inverse function
LNOR/LNORMAL - lognormal
LNRN - Lognormal random number
loc - location
LOG1 - one-parameter logistic (Rasch)
LOG2 - two-parameter logistic
LOGIT - logistic regression
LOGITHYPO - hypothesis tests in logistic regression
LOGLIN - loglinear modeling
LR - likelihood ratio
LRCHI - likelihood ratio chi-square
LRDEV - likelihood ratio of deviate
LRN - Logistic random number
LS - least-squares
LSD - least significant difference
LSL - lower specification limit
LSQ - least-squares
LTAB - life tables
LTL - lower tolerance limit
LW - Lawless and Wang

M
MA - moving average

MAD - mean absolute deviation
MAHAL - Mahalanobis distances
MANCOVA - multivariate analysis of covariance
MANOVA - multivariate analysis of variance
MANOVAHYPO - hypothesis tests in MANOVA
MANOVAPOST - post hoc estimate for repeated measures in MANOVA
MAR - missing at random
MAX - maximum
MAXSTEP - maximum number of steps
MCAR - missing completely at random
MCMC - Markov Chain Monte Carlo
MDPREF - multidimensional preference
MDS - multidimensional scaling
MIN - minimum
M-H- Metropolis-Hastings
MIS - number of missing values
MIX - mixed regression
MIXHIER - mixed regression for data having a hierarchical structure
MIXMULTY - mixed regression for data having a multivariate structure
ML - Maximum Likelihood
MLA - maximum likelihood analysis
MLE - maximum likelihood estimate
MML - maximum marginal likelihood
MRC - Multiple Regression and Correlation
MS - mean squares
MSE - mean square error
MSIGMA - sigma measurement
MT - Mersenne-Twister
MTW - MINITAB v11 data files
MU2 - Guttman's mu2 monotonicity coefficients
MULTIVAR - multiple variables
MW - minimum within sum of squares deviations
MWL - maximum Wishart likelihood

N

NAR - non-stationary first-order autoregressive
NB - nominal the best
NBB - nominal-the-best: bilateral tolerance
NBCF - Negative binomial cumulative function
NBD - number of active bounds on parameter values
NBDF - Negative binomial density function
NBIF - Negative binomial inverse function
NBINOMIAL - Negative binomial
NBRN - Negative binomial random number
NBU - nominal-the-best: unilateral tolerance
NCAT - number of categories
NCF - Binomial cumulative function
NCOL - number of columns
NDF - Binomial density function
NDMAX - maximum number of points
NDMIN - minimum number of points
NEM - number of EM iterations
NEXPO - negative exponential
NIF - Binomial inverse function
NIPALS - Nonlinear iterative partial least Squares
NLAG - number of lags
NLLOSS - nonlinear loss functions
NLMODEL - nonlinear models
NMIN - minimum count
NMULTIVAR - multiple numeric variables
NONLIN - nonlinear models
NP-Number nonconforming
NPAR - nonparametric
NREC - non-recreationist
NRN - Binomial random number
NROW - number of rows
NRP - number of apparently redundant parameters
NSAMP - number of sub-samples
NSPLIT - maximum number of splits
NX - number of nodes along the x axis
NXDIS - number of discretization points in the x (North) direction
NY - number of nodes along the y axis
NYDIS - number of discretization points in the y (East) direction
NZ - number of nodes along the z axis

NZDIS - number of discretization points in the z (Depth) direction

O

Obs-observed
OBSFREQ - observed frequency
OC - operating characteristic
ODBC - open database capture and connectivity
OFREQ - outlier frequencies
OLS - ordinary least-squares
ORTHEQ-Equally Spaced Orthogonal component
ORTHUN- Unequally Spaced Orthogonal component

P

P - Proportion nonconforming
PACF - Pareto cumulative function
PACF - partial autocorrelation function
PADF - Pareto density function
PAIF - Pareto inverse function
PARAM - parameters
PARN - Pareto random number
PCA - process capability analysis
PCF - iterated principal axis factoring
PCF - Poisson cumulative function
PCNTCHANGE - percentage change
PCT - Macintosh PICT
PDF - Poisson density function
pdf - probability density function
PDL - polynomial distributed lag
PERMAP - perceptual mapping
PIF - Poisson inverse function
PLIMITS - probability limits
PLS - partial least squres
pmf - probability mass function
PMIN - minimum proportion
PNG - Portable Network Graphics
POLY - polygon
POSAC - partially ordered scalogram analysis with coordinates
P-P - probability plot
PP - process performance
Ppk - Process performance index for off-centered process
PPL - process performance based on lower specification limit
PPM - parts per million
PPU - process performance based on upper specification limit
PRE - percentage reduction error
PREFMAP - preference mapping
PRN - Poisson random number
PROB - probability
PROP1 - single proportion
PROP2 - equality of two proportions
PS - PostScript
PVAF/p.v.a.f. -- present value annuity factor
p-value - probability value

Q

QC - quality control
QMLE - quasi maximum likelihood estimate
QNTL - quantiles
QPLOT - quantile plots
Q-QPLOT - two sample quantile plot
QRD - QR decomposition
QS - quick search
QSK - quantitative symmetric similarity coefficients (or Kulczynski measure)
QUASI - Quasi-Newton method

R

R & R - repeatability and reproducibility
R chart - range chart
RADMAX - maximum horizontal direction for the search radius
RADMIN - minimum horizontal direction for the search radius
RAND - random
RANDSAMP - random sampling
RANKREG - rank regression

RBSTAT - row basic statistics
RCF - Rayleigh cumulative function
RDF - Rayleigh density function
RDISCRIM - robust discriminant
RDIST - robust distance
RDVER - vertical direction for the search radius
REPAR - reparametrize
REPS - replicates
RESID - residuals
RIF - Rayleigh inverse function
RJS - rejection sampling
RMS - root mean square
RMSEA - root mean square error of approximation
RMSSTD - root mean square standard deviation
ROC - receiver operating characteristic
ROWPCT - Row percentages
RRN - Rayleigh random number
RS - response surface
RSE- robust standard errors
RSEED - random seed
RSM- response surface methods
RSQ - stress and squared correlation
RSS - residual sum of squares
RSTATISTICS - row statistics
RTF - rich text format
RWM-H - random walk Metropolis-Hastings
RWSTEM - stem and leaf plot for rows

S

S chart - standard deviation control chart
SANG1 - angle (in degrees) of the first minor axis of the search ellipsoid
SANG2 - angle (in degrees) of the major axis of the search ellipsoid
SANG3 - angle (in degrees) of the second minor axis of the search ellipsoid
SAV - SPSS files
SB - smaller the better
sc - scale
SC - set correlation
SCDFUNC - standardized coefficients for canonical variables
SCF - Studentized cumulative function
SD - standard deviations
sd2/sas7bdat - SAS v9 files
SDF - Studentized density function
SE/se/S.E. - standard error
SEK - standard error of kurtosis
SEM - standard error of mean
SES - standard error of skewness
shp - shape
SIF - Studentized inverse function
SIMPLS - Straight-forward Implementation of Partial Least Squares
SKMEAN - simple kriging mean
SL - specification limit
SMIN - minimum split value
SPLOM - scatter plot matrix
SQL - structured query language
SQRT/SQR - square-root
SRN - Studentized random number
SRWR - sum of rank weighted residuals
SS - sum of squares
SSCP - sum of squares and cross products
STA - Statistica v5 data files
STAND - standardized deviates
SVD - singular value decomposition
SW - Shapiro-Wilks
SYC/CMD - SYSTAT command Files
SYZ/SYD/SYS - SYSTAT data files
SYO - SYSTAT output files

T

T1 - one-sample t-test
T2 - two-sample t-test
TANALYZE - Taguchi design: analyze
TCF - t cumulative function
TCOR - total correlation
TCOV - total covariance
TDF -t density function
TESTAT - Test Item Analysis

TESTATCL - classical test item analysis
TESTATLOG - logistic item response analysis
TETRA - tetrachoric correlations
TGENERATE - Taguchi design: generate
TIF - t inverse function
TIFF - Tagged Image File Format
TLOG - log time
TLOSS - Taguchi's Loss Function
TNH - hyperbolic tangent
TOHC0 - Hypothesis Testing: Zero correlation
TOHC1 - Hypothesis Testing: Specific correlation
TOHC2 - Hypothesis Testing: Equality of two correlation coefficients
TOHP1 - Hypothesis Testing: Single proportion
TOHP2 - Hypothesis Testing: Equality of two proportions
TOHT1 - Hypothesis Testing: One sample t-test
TOHT2 - Hypothesis Testing: Two sample t-test
TOHTPAIRED - Hypothesis Testing: Paired t-test
TOHV1 - Hypothesis Testing: Single variance
TOHV2 - Hypothesis Testing: Two variances
TOHVN - Hypothesis Testing: Several variances
TOHZ1 - Hypothesis Testing: One sample z-test
TOHZ2 - Hypothesis Testing: Two sample z-test
TOL - tolerance
TPLOT - time series plot
TPREDICT - Taguchi design: predict
TRCF - Triangular cumulative function
TRDF - Triangular density function
TRI - triangular
TRIF - Triangular inverse function
TRIM - trimmed mean
TRN - t random number
TRP - transpose
TRRN - Triangular random number
TSFOURIER - Fourier decomposition of time series
TSIV - Two-Stage Instrumental Variables
TSLS - Two-Stage Least Squares
TSP - traveling salesman path
TSQ chart - Hotelling's T^2 chart
TSSMOOTH - smoothing time series
TXT - text format

U
U chart - chart showing defects per unit
UCF - Uniform cumulative function
UCL - upper control limit
UDF - Uniform density function
UIF - Uniform inverse function
UNCE - uncertainty coefficient
URN - Uniform random number
USL - upper specification limit
UTL - upper tolerance limit

V
VAR - variance
VIF - variance inflation factor

W
WB - Weibull
WCF - Weibull cumulative function
WCOR - pooled within-group correlation
WCOV - pooled within-group covariance
WDF - Weibull density function
WHISKER - Box-and-Whisker plot
WIF - Weibull inverse function
WMF - Windows metafile
WRN - Weibull random number

X
XCF - Chi-square cumulative function
XDF - Chi-square density function
XIF - Chi-square inverse function
XLAG - separation distance between lags
XLS - excel format
XLTOL - tolerance for lags
XMAX - maximum along x axis
XMIN - minimum along x axis

X-MR chart - Individuals and moving range chart
XPT/TPT - SAS transport files
XRN - Chi-square random number
XTAB - Crosstabulations

Y
YMAX - maximum along y axis
YMIN - minimum along y axis

Z
Z1 - one-sample z-test
Z2 - two-sample z-test
ZCF - Normal cumulative function
ZDF - Normal density function
ZICF - Zipf cumulative function
ZIDF - Zipf density function
ZIF - Normal inverse function
ZIIF - Zipf inverse function
ZIRN - Zipf random number
ZMAX - maximum along z axis
ZMIN - minimum along z axis
ZRN - Normal random number

Index

A

B

C

G

M

T

U

V

W

X

Y

Z